# FEATURE MINING WITH COMPUTATIONAL INTELLIGENCE AND ITS APPLICATIONS IN IMAGE STEGANALYSIS AND BIOINFORMATICS

By

Qingzhong Liu

Submitted in Partial Fulfillment
of the Requirements for the

Doctorate of Philosophy in Computer Science

New Mexico Institute of Mining and Technology
Department of Computer Science

Socorro, New Mexico
July 2007

# ACKNOWLEDGEMENTS

Many people contributed to the success of this work. First, I gratefully acknowledge my advisor, Dr. Andrew Sung, whose advice, guidance, passion, and support of this study was indispensable. Dr. Sung has been instructing my Ph.D. study since I joined the Computer Science Department of New Mexico Institute of Mining and Technology (NMT) in 2002. He is a great mentor with broad and long view and continuous scientific passion. I truly appreciate his advice and guidance in my research and his generous support throughout my PhD study as well as his encouraging and affording me to attend the prestigious international academic conferences, and hence broaden my horizon and make me thriving. Without him, I could not have this opportunity of pursuing my Ph.D. at the NMT of U.S.A., let alone the finish of this thesis and my growth up in the past five years.

I would like to thank the other members of my committee, Dr. Bernardete Ribeiro, Dr. Srinivas Mukkamala, and Dr. Dongwan Shin, who have provided assistance and guidance during my PhD study. They also gave me helpful suggestions on my study. I am very grateful for their help.

Appreciations go to Dr. Soliman, Dr. Liebrock, Dr. Mazamdar, and Dr. Clausen, who taught me the knowledge of several subjects in computer science. I got much expertise and knowledge from them since I was their Teaching Assistant.

I thank Drs. Lyu, Simoncelli and Harmsen for providing me their codes which are used in the comparison study in steganalysis (chapters 2, 3, and 4) and Drs. Masotti, Tang, Zhou, and Diaz-Uriarte for sending me or publicizing their codes which are used in the comparison study in Microarray analysis (chapter 6). Special thank goes to Dr. Liang for her insightful discussion and offering me the SNP data sets which are tested in chapter 7.

The friendship of Dr. Jianyun Xu and Dr. Zhongxue Chen is highly appreciated and has led to many interesting and good-spirited discussions related to this research.

Without a shadow of doubt, my English tutors Lisa, Kay, and Jonathan did very good job in helping me to improve my English. I appreciate them for their very helpful tutoring, great patience, precious time, and love from the bottom of my heart.

Last, but not least, I would acknowledge my family for their understanding and love during the past years. Their support, encouragement, passion, and love were in the end what made this dissertation possible.

הַלְלוּיָהּ ( Hallelujah)!

# ABSTRACT

Steganalysis aims to detect the information-hiding behavior in steganographic systems. Bioinformatics is to solve the biological problems usually on the molecular level. Although steganalysis and bioinformatics are completely different, both of them involve feature mining and computational intelligence techniques. It is very challenging to solve the problems in these two fields. In this thesis, chapter 1 is the introduction on steganography and steganalysis and chapter 5 is the introduction on bioinformatics; the contributions in image steganalysis are presented in chapters 2, 3, and 4 and the contributions in bioinformatics are presented in chapters 6 and 7, described as follows.

Information-hiding ratio is a well-known reference to evaluation of the detection performance in steganalysis. In chapter 2, I introduce another parameter of image complexity to evaluation of the performance, and present a scheme of steganalysis of Least Significant Bit (LSB) matching steganography based on feature extraction and pattern recognition techniques. Comparing to other well-known methods of steganalysis of LSB matching steganography, our method performs the best. The significance of features and the detection performance depend not only on the information-hiding ratio but also on the image complexity.

In chapter 3, I present a scheme based on feature mining and pattern classification to detect LSB matching steganography in grayscale images, which is a very challenging problem in steganalysis. Different types of features are proposed. In comparison with other well-known feature sets, the set of proposed features performs the best. I compare

different learning classifiers and deal with the issue of feature selection that is rarely mentioned in steganalysis. In our experiments, the combination of a Dynamic Evolving Neural Fuzzy Inference System (DENFIS) with a feature selection of Support Vector Machine Recursive Feature Elimination (SVMRFE) achieves the best detection performance. Results also show that image complexity is an important reference to evaluation of steganalysis performance.

Based on the Generalized Gaussian Distribution (GGD) model in the quantized DCT coefficients, the errors between the logarithmic domain of the histogram of the DCT coefficients and the polynomial fitting are extracted as features to detect the adulterated JPEG images and the untouched ones. Computational intelligence techniques are applied to extracted features. The designed method is successful in detecting the information-hiding types and the information-hiding length in the multi-class JPEG images including the CryptoBola, F5, and JPHS steganographic systems. The details are described in chapter 4.

Chapter 6 aims to improve the classification of microarray gene expression data, which have a high dimension of variables and small sample size. Gene selection is very important to the classification. Most existing gene selection methods, including modified test statistic-based approaches and model-based approaches such as logistic model or mixed models, give highly correlated significant genes that are redundant for classification. I develop a new gene selection method, Recursive Feature Addition (RFA), which combines supervised learning and statistical measures for the chosen candidate

genes to deal with the redundant information. I also propose an algorithm of Lagging Prediction Peephole Optimization (LPPO) to choose the final feature set.

Comprehensive evaluation of common genetic variations through association of SNP structure with common complex diseases in the genome-wide scale is currently a hot area in human genome research. Exploiting information redundancy due to associations between single nucleotide polymorphism (SNP) markers potentially reduces the efforts in terms of time and cost for these studies. One of the fundamental questions in SNP-disease association study is how many SNPs is enough to provide good prediction performance of disease status. In chapter 7, I develop a new feature selection method named Supervised Recursive Feature Addition (SRFA). This method combines supervised learning and statistical measures for the chosen candidate features/SNPs to deal with the redundancy information so that it can improve the classification in association studies. Additionally, I propose a Support Vector based lowest weight and lowest correlation Recursive Feature Addition (SVFRA) scheme in SNP-diseases association analysis. Results show that on the average, our SRFA outperforms the well-known method of Support Vector Machine Recursive Feature Elimination and logic regression based SNP selections for disease classification in genetic association study.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION TO STEGANOGRAPHY AND STEGANALYSIS

Steganography is the art and science of communicating hidden messages in such a way that no one apart from the intended recipient knows of the existence of the message; this is in contrast to cryptography, where the existence of the message itself is not disguised, but the content is obscured. The word "*Steganography*" is of Greek origin and means *"covered, or hidden writing"*. Its ancient origins can be traced back to 440 BC. Herodotus mentions two examples of Steganography in *The Histories of Herodotus* [Petitcolas *et al.* 1999]. Demeratus sent a warning about a forthcoming attack to Greece by writing it on a wooden panel and covering it in wax. Wax tablets were in common use then as re-usable writing surface, sometimes used for shorthand. Another ancient example is that of Histiaeus, who shaved the head of his most trusted slave and tattooed a message on it. After his hair had grown the message was hidden. The purpose was to instigate a revolt against the Persians. Later, Johannes Trithemius's book *Steganographia* is a treatise on cryptography and steganography disguised as a book on black magic.

The advantage of using steganography over using cryptography alone is that the secret messages will not attract attention. An unhidden coded message, no matter how unbreakable it is, will arouse suspicion. Generally, we can hide data in digital media including images, audios, and videos as well as TCP/IP packets, etc. Currently, digital image is one of the most popular media types for carrying covert message. The innocent image is called carrier or cover; and the adulterated image carrying some hidden message is called stego-image or steganogram. Fig. 1-1(a) is an example of steganogram wherein the text-file about Alzheimer's disease is hidden. Fig. 1-1(b) lists the covert texts.

(a)

Alzheimer's: The Mysteries of the Most Common Form of Dementia

In November of nineteen ninety-four, Ronald Reagan wrote a letter to the American people. The former president shared the news that he had Alzheimer's disease. Mister Reagan began what he called his journey into the sunset of his life. That ten year journey ended on June fifth, two thousand four, at the age of ninety-three.

In his letter, America's fortieth President wrote about the fears and difficulties presented by Alzheimer's disease. He said that he and his wife Nancy hoped their public announcement would lead to greater understanding of the condition among individuals and families affected by it.

Ronald Reagan was probably the most famous person to suffer from Alzheimer's disease. In the United States, about four million five hundred thousand people have the disease. Many millions more are expected to have it in years to come.

Doctors describe Alzheimer's as a slowly increasing brain disorder. It affects memory and personality -- those qualities that make a person an individual. There is no known cure. Victims slowly lose their abilities to deal with everyday life. At first they forget simple things, like where they put something or a person's name.

As time passes, they forget more and more. They forget the names of their husband, wife or children. Then they forget who they are. Finally, they remember nothing. It is as if their brain dies before the other parts of the body. Victims of Alzheimer's do die from the disease, but it may take many years.

(b)

Fig. 1-1 An example of steganogram. The covert message shown in (b) is embedded in the left image (a).

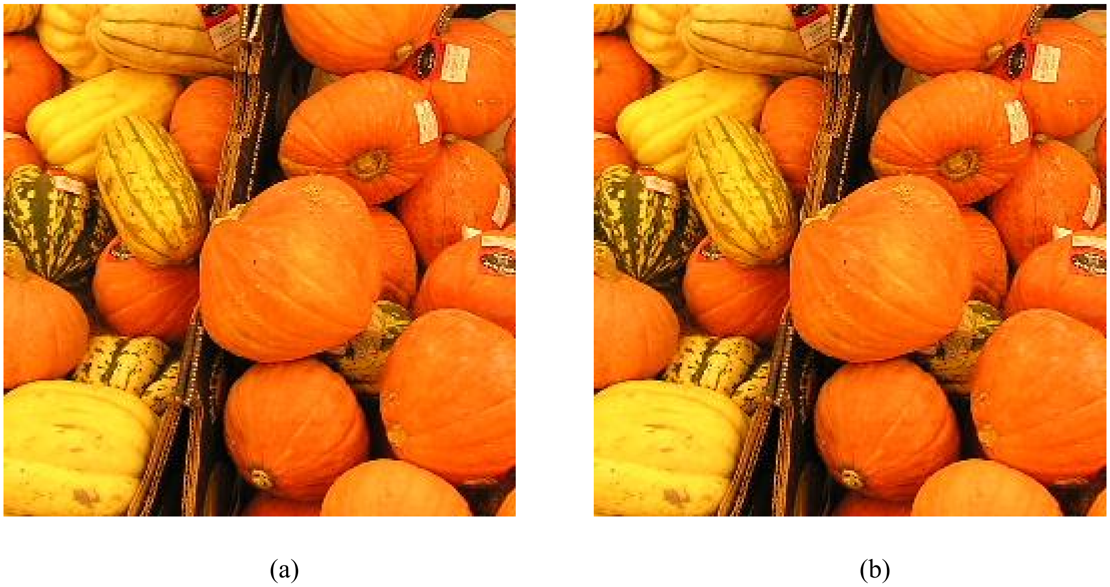
Though not proven, there have been claims that terrorists have been using steganography to communicate with each other in planning attacks. It has been thought that images with embedded messages have been placed on bulletin boards or dead drops for other terrorists to pick up and then retrieve any hidden messages. Since it is so difficult to detect when steganography is taking place, this is a very secure form of communication and it has thought to be used by Al-Qaida [Kelley 2001, http://www.usatoday.com/tech/news/2001-02-05-binladen.htm].

The common information-hiding techniques implement hiding data in digital images by modifying the pixel values of the space domain or modifying the transform coefficients.

In hiding data in the space domain, one simple method is Least Significant Bit (LSB) steganography or LSB embedding [Kurak and McHugh, 1992]. Each byte of an image represents a different color. The last few bits in a color byte, however, do not hold as much significance as the first few. Therefore, two bytes that only differ in the last few bits can represent two colors that are virtually indistinguishable to the human eye. For example, 00100100 and 00100101 can be two different shades of red, but since it is only the last bit that is different, it is impossible to see the color difference. LSB embedding, then, alters these last bits by hiding a message within them. LSB embedding has the merit of simplicity, but suffers from the lack of robustness. LSB matching, another method of hiding data in space domain randomly alters the bytes by plus or minus one according to the bit of cipher message, not simply replacing the last bits [Sharp, 2001].

In hiding data in the transform domain, a message is embedded by the way of modifying transform coefficients of the cover. There are three common transform techniques: Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT). For example, hiding data in the low frequency part of 2-D lossless wavelet transform and utilizing convolution error correction coding, Xu *et al.* designed different information-hiding systems by embedding data in the wavelet domain to achieve a big hiding capacity and extremely robustness against JPEG compression [Xu *et al*., 2003, 2004]. Derek Upham publicized JPEG-JSteg to hide data in JPEG images [Derek Upham, http://www.nic.funet.fi/pub/crypt/steganography]. Its embedding algorithm sequentially replaces the least-significant bit of DCT coefficients with the message's data

[Provos and Honeyman, 2003]. And it is easy to be detected [Zhang and Ping, 2003]. Instead of replacing the least-significant bit of DCT coefficient with message data, F5 decrements its absolute value in a process called matrix encoding [Westfeld, 2001]. Additionally an efficient FFT based signal scheme for multimedia steganography was proposed to permit the use of signal sets of large dimensions without increasing the computational complexity drastically [Ramkumar *et al.*, 1999].



(a) (b)

Fig. 1-2 An original cover (a) and the stego-image (b)

Other information hiding techniques include spread spectrum steganography [Marvel *et al.*, 1999], statistical steganography, distortion, and cover generation steganography [Katzenbeisser and Petitcolas, 2000], etc. Many hiding tools can be downloaded from Internet based on different hiding methods such as Invisible Secrets [http://www.invisiblesecrets.com], Secure Engine [http://securengine.isecurelabs.com/, retrieved on Apr 27, 2007], Hide4PGP [http://www.heinz-repp.onlinehome.de/Hide4PGP.htm], and CryptoBola [http://www.cryptobola.com]. Fig. 1-2 shows a JPEG image (a) and the steganogram (b) wherein 628-byte data is hidden. It is very challenging to judge which one is carrying the hidden data in the existence of both the cover and the stego-image, let alone in the single appearance of the cover or the stego-image.

Steganalysis aims to discover the presence of hidden data. Westfeld performed the blind steganalysis on the basis of statistical analysis of PoVs (pair of values). This method, so-called $\chi^2$-statistical analysis [Westfeld and Pfitzmann, 2000], gave a successful result to a sequential LSB (Least Significant Bit) embedding steganography. Provos extended this method by re-sampling the test interval and re-pairing values [Provos, 2001]. Fridrich *et al.* introduced a RS steganalysis which is based on the partition of an image's pixels into three groups: Regular, Singular and Unusable and estimate the possible embedded message length of the LSB steganography [Fridrich, Goljan and Du, 2001]. Lyu and Farid [Lyu and Farid, 2004, 2005] described an approach to detect hidden messages in images that uses a wavelet-like decomposition to build higher-order statistical models of natural images. Support vector machines are then used to discriminate between untouched and adulterated images. Avcibas *et al.* presented a universal detection technique for steganalysis of image based on image quality metrics [Avcibas *et al*., 2003]. Based on the 3-D DFT and the calculation of the center of mass, Harmsen and Pearlman proposed a detector of the Histogram Characteristic Function Center Of Mass (HCFCOM) that is very successful in detecting multiple information-hiding systems [Harmsen and Pearlman, 2003]. Based on HCFCOM, Ker designed Adjacency HCFCOM (A.

HCFCOM) and Calibrated Adjacent HCFCOM (C. A. HCFCOM) to improve the probability of detection for LSB matching in grayscale images [Ker, 2005].

To this date, most publications refer information hiding ratio to evaluate the performance of steganalysis. Specifically, the higher the hiding ratio, the higher the detection performance will be. However, to our knowledge, few publications mentioned the parameter of image complexity that is also very important to evaluate the detection performance. On the other side, most of steganalysis methods depend on feature design and pattern classification techniques. But the feature selection was rarely mentioned in the past literatures in steganalysis. These two issues will be addressed as well as new detection methods designed in the following chapters.

The remainder of steganalysis is organized as follows. Chapter 2 describes the steganalysis of LSB matching steganography which is one of the most difficult space-hiding steganography for detection, and introduces the shape parameter of Generalized Gaussian Distribution (GGD) in the wavelet domain to measure the image complexity and evaluate the steganalysis performance. Chapter 3 presents new features to improve the detection performance in stgeganalysis of LSB matching steganography in grayscale images and deals with the feature selection in the steganlysis. Chapter 4 is the detection of the information-hiding behavior in transform-hiding steganography, focusing on JPEG images.

# CHAPTER 2: STEGANALYSIS OF SPACE-HIDING STEGANOGRAPHIC SYSTEMS

## 2.1 Introduction

Space-hiding steganographic system implements embedding data in the space domain. Specifically, for image, it modifies the pixel values to achieve the goal of hiding data. A popular information-hiding technique in space-hiding steganographic systems in images is Least Significant Bit (LSB) Embedding/Replacement, which combines high capacity with visual imperceptibility and very ease of implementation. However this information-hiding system has the weakness to the sensitive statistical detections such as $\chi^2$-test and RS-steganalysis. A minor modification of the LSB Embedding/Replacement method, which we call LSB Matching, retains the favorable characteristics of LSB Replacement but is more difficult to detect statistically.

## 2.2 LSB Matching and Related Work on the Detection

LSB Matching was first described by Sharp [Sharp, 2001]. ]. In each case the secret data is taken as a stream of bits, and the cover image is considered as a stream of bytes. These bytes are taken in a pseudorandom order, as specified by a secret key which is presumed to be shared between sender and recipient of the stego image. This serves both to prevent the enemy steganalyst from reading the secret data straight off and also to spread the secret data over the cover when there is less than the maximal amount. Yu *et al.* designed the LSB matching steganography evading the statistical analyses of $\chi^2$-test and RS-steganalysis [Yu *et al*., 2004]. The idea is to preserve the occurrence of PoVs by applying the random flipping to embedding a message and to adjust the RS statistical measures with unused embedding parts after embedding a secret message. Since LSB matching is hard for detection and easy for implementation, it's important and challenging to design a reliable method for detecting the information-hiding.

There are a few detectors that may be used in detecting the information-hiding in LSB matching steganography. One of them is the histogram characteristic function center of mass (HCFCOM) [Harmsen and Pearlman, 2003] since the embedding of LSB matching can be modeled on noise adding. To improve the probability of detection for LSB matching in grayscale images, based on the Harmsen and Pearlman's contribution, Ker proposed Adjacency HCFCOM (A. HCFCOM) and Calibrated Adjacency HCFCOM (C. A. HCFCOM) [Ker, 2005]. Farid and Lyu achieved an approach to detecting hidden messages in images by using a wavelet-like decomposition to build high-order statistical models of natural images [Lyu and Farid, 2004, 2005]. Fridrich *et al.* presented a Maximum Likelihood (ML) estimator for predicting the hiding ratio of non-adaptive ±K embedding in images [Fridrich *et al*., 2005]. Unfortunately, the ML estimator starts to "fail to reliably estimate the message length once the variance of the sample exceeds 9" [Fridrich *et al*., 2005]. Recently, correlation features in spatial domain and wavelet domain are extracted for image steganalysis [Liu, Sung and Ribeiro, 2005], although the method is effective for detection of several steganography systems, the images in the experiments are downloaded from Internet and the almost all of them are compressed. It is not done on the experiments on never compressed images. Generally, regarding the information-hiding in space domain, it is much more difficult to detect the information-hiding in never compressed images than that in compressed images.

To our knowledge, most publications evaluate the steganalysis performance in reference to information hiding ratio and miss another important parameter of image complexity that is also very important in evaluating the detection performance. In this chapter, the shape parameter of the Generalized Gaussian Distribution (GGD) in the wavelet domain is introduced to measure the image complexity, as a reference as well as the information-hiding ratio to the evaluation of the steganalysis performance. Different types of features are designed for detection of the information-hiding in LSB matching steganography.

## 2.3 Image Complexity and GGD model

Several articles [Huang and Mumford, 1999; Sharifi and Leon-Garcia, 1995; Wainwright and Simoncelli, 2000; Wouwer *et al.*, 1999; Winkler, 1995] describe the statistical models of images such as Markov Random Field models (MRFs), Gaussian Mixture Model (GMM), and Generalized Gaussian Distribution (GGD) model in transform domains, such as, DCT, DWT, and Discrete Fourier Transform (DFT).

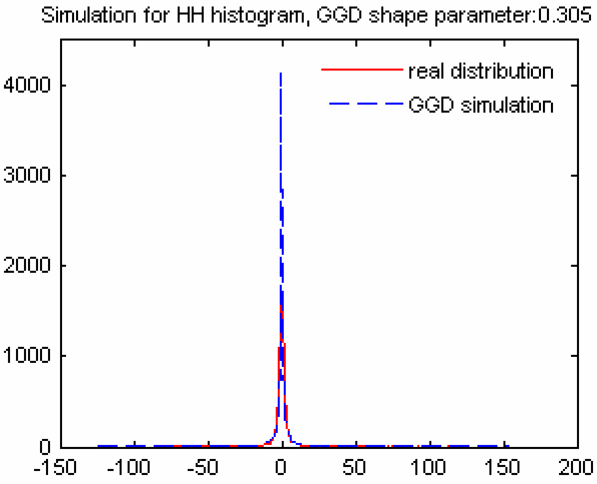
Experiments show that a good Probability Distribution Function (PDF) approximation for the marginal density of coefficients at a particular sub-band produced by various types of wavelet transforms may be achieved by adaptively varying two parameters of the GGD [Sharifi and Leon-Garcia, 1995; Moulin and Liu, 1999], which is defined as

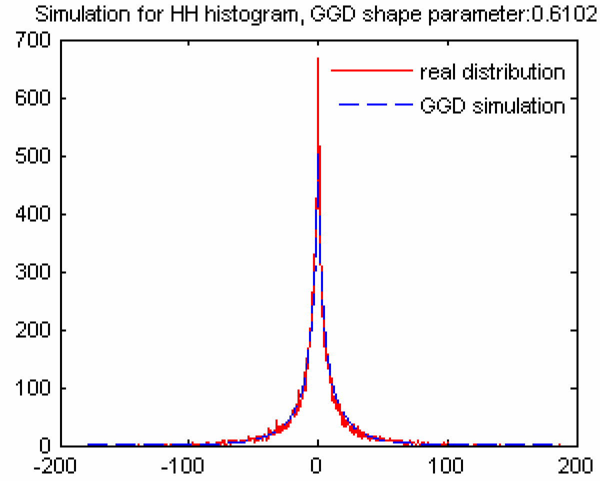
$$p(x;\alpha,\beta)=\frac{\beta}{2\alpha\Gamma(1/\beta)}e^{-(|x|/\alpha)^{\beta}} \tag{2-1}$$

where $\Gamma(z)=\int_0^{\infty}e^{-t}t^{z-1}dt, z>0$ is the Gamma function, the scale parameter *α* models the width of the PDF peak (standard deviation), and the shape parameter *β* is inversely proportional to the decreasing rate of the peak. The GGD model contains the Gaussian and Laplacian PDFs as special cases, using *β* = 2 and *β* = 1, respectively.

Generally, an image with high complexity has a high shape parameter to the GGD in the wavelet domain. Fig. 2-1 shows the 256×256 grayscale images with different textures on the left and the histogram distributions of the Haar wavelet HH sub-band coefficients and the GGD simulations on the right.

The fact that the high peak distribution of the wavelet coefficients is obtained at the value of zero indicates that adjacent pixels are highly correlated. More clearly, Fig. 2-2(a) shows an 8-bit grayscale image. The variable *v*(*i*, *j*) denotes the grayscale value at point (*i*, *j*) and *v*(*i*+1, *j*) denotes the grayscale value at the point (*i*+1, *j*). The occurrences of the pair (*v*(*i*, *j*), *v*(*i*+1, *j*)) calculate the joint distribution of the adjacent points, shown in Fig. 4(b) which demonstrates the high correlation of adjacent pixels.


Simulation for HH histogram, GGD shape parameter:0.305
real distribution
GGD simulation
4000
3000
2000
1000
0
-150
-100
-50
0
50
100
150
200


Simulation for HH histogram, GGD shape parameter:0.6102
real distribution
GGD simulation
700
600
500
400
300
200
100
0
-200
-100
0
100
200

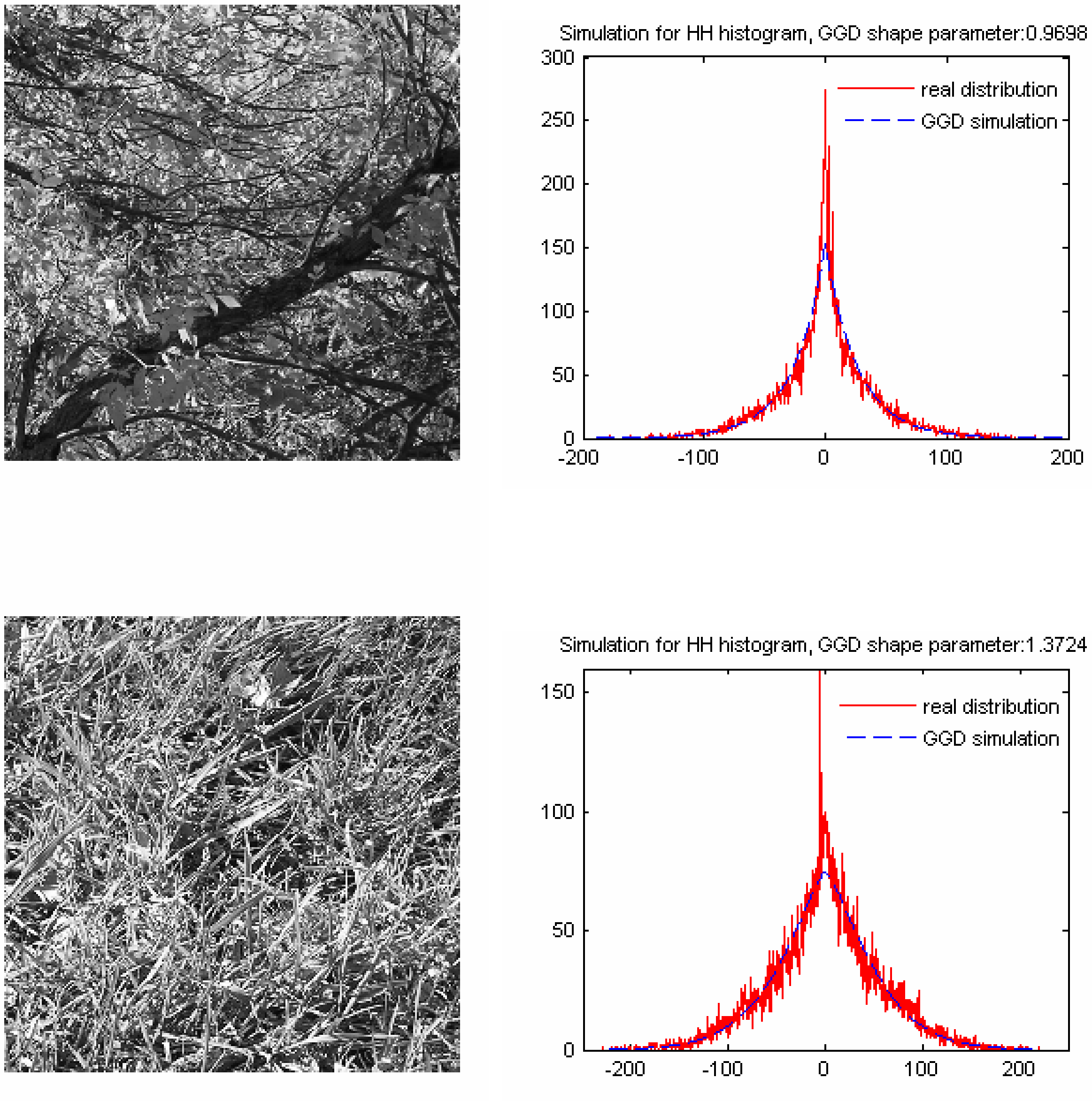
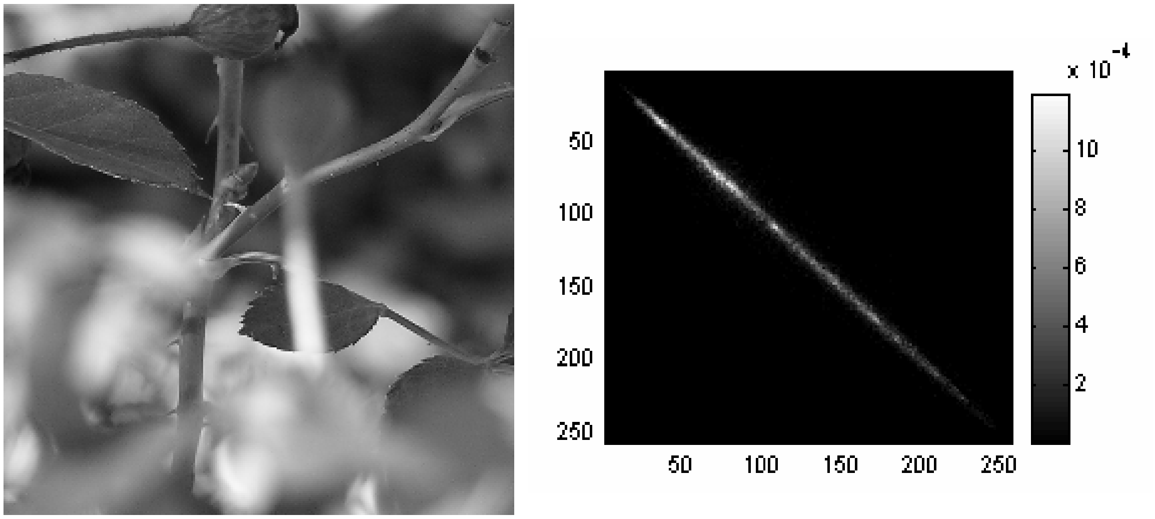


Fig. 2-1 Demonstration of image complexity and the GGD. The 256×256 grayscale images with different complexity (left) and the generalized Gaussian distribution of the HH sub-band coefficients (right), decomposed by Haar wavelet. Fig. 2-1 indicates that the image with low complexity has low shape parameter of the GGD and the image with high complexity has high shape parameter of the GGD.



(a) (b)

Fig. 2-2 An 8-bit grayscale image (a) and the joint distribution (b) of the adjacent pixel pair (v(i, j), v(i+1, j)). The horizontal axis in (b) shows the value of the pixel (i, j) and the vertical axis marks the value of the pixel (i +1, j). The joint distribution indicates the probability of the pair (v(i, j), v(i+1, j)).

## 2.4 Feature Extraction

Based on the GGD model and the observation of the high correlation of the adjacent pixels, mentioned in 2.3, here is the hypothesis that the information-hiding in space-hiding systems will affect the high correlation of the adjacent pixels. Based on this hypothesis, the following features are designed.

We consider the correlation between LSBP and the second Least Significant Bit Plane (LSBP2). $M_1 = \{b_1^{ij}\}$($i$=1, 2, …, $m$; $j$ = 1, 2, …, $n$; $i$ and $j$ give the location of the element in the matrix) is the $m$×$n$ matrix of the binary bits of the LSBP and $M_2 = \{b_2^{ij}\}$($i$=1, 2, …, $m$; $j$ = 1, 2, …, $n$) is the $m$×$n$ matrix of binary bits of the LSBP2. Here $m$ and $n$ are the numbers of pixels in horizontal and vertical directions, and $E(\bullet)$ is the mathematical expectation. The covariance function is defined as

$$Cov(x_1, x_2) = E[(x_1 - u_1)(x_2 - u_2)] \tag{2-2}$$

where $u_i = E(x_i)$.

$C1$ is defined as follows:

$$C1 = cor(M_1, M_2) = \frac{Cov(M_1, M_2)}{\sigma_{M_1}\sigma_{M_2}} \tag{2-3}$$

where $\sigma_{M_1}^2 = Var(M_1)$ and $\sigma_{M_2}^2 = Var(M_2)$.

The autocorrelation $C$(k, l) of LSBP is defined as follows:

$$C(k,l) = cor(X_{11(m-k)(n-l)}, X_{(k+1)(l+1)mn}) \tag{2-4}$$

where, $X_{11(m-k)(n-l)} = \{b_1^{ij}\}(i = 1, 2, ..., m-k; j = 1, 2, ..., n-l)$

and $X_{(k+1)(l+1)mn} = \{b_1^{ij}\}(i = k+1, k+2, ..., m; j = l+1, l+2, ..., n)$.

Set different values to *k* and *l*, the features from *C*2 to *C*15 are presented as follows:

$C2 = C(1, 0)$; $C3 = C(2, 0)$; $C4 = C(3, 0)$; $C5 = C(4, 0)$;

$C6 = C(0, 1)$; $C7 = C(0, 2)$; $C8 = C(0, 3)$; $C9 = C(0, 4)$;

$C10 = C(1, 1)$; $C11 = C(2, 2)$; $C12 = C(3, 3)$; $C13 = C(4, 4)$;

$C14 = C(1, 2)$; $C15 = C(2, 1)$.

The variable $\rho_k$ denotes the histogram probability density of cover at the intensity, *k* (*k* = 0,1, …,N-1, for 8-bit grayscale image, N = 256). The variable, $\rho'_k$, denotes the histogram probability density of adulterated images at the intensity *k*. The LSBP hiding rate *r* is the relative length of hidden binary data, assume the hidden data is i.i.d., for 8-bit grayscale image, $\rho'_k$ is given as follows:

$$\rho'_k = (1\text{-r}/2)^* \rho_k + (\text{r}/4)^* \rho_{k\text{-}1} + (\text{r}/4)^* \rho_{k+1}, \quad 2 \le k \le 253$$

$$\rho'_0 = (1\text{-r}/2)^* \rho_0 + (\text{r}/4)^* \rho_1$$

$$\rho'_1 = (1\text{-r}/2)^* \rho_1 + (\text{r}/4)^* \rho_2 + (\text{r}/2)^* \rho_0 \qquad (2\text{-}5)$$

$$\rho'_{255} = (1\text{-r}/2)^* \rho_{255} + (\text{r}/4)^* \rho_{254}$$

$$\rho'_{254} = (1\text{-r}/2)^* \rho_{254} + (\text{r}/4)^* \rho_{253} + (\text{r}/2)^* \rho_{255}$$

Without original cover, just based on the distribution density of the histogram, it is too difficult to accurately judge that the test image is hiding some data or not and predict the hiding ratio *r*. However, LSB matching steganography definitely modifies the distribution density of the histogram. Based on this point, we present the correlation features on the histogram. The histogram probability density, *H*, is denoted as $(\rho_0, \rho_1, \rho_{2,} \ldots, \rho_{\text{N-1}})$. The histogram probability densities, $H_e$, $H_{\text{o}}$, $H_{l1,}$ and $H_{l2}$ are given:

$$H_e = (\rho_0, \rho_2, \rho_4 \ldots \rho_{\text{N-2}}) , \qquad H_o = (\rho_1, \rho_3, \rho_5 \ldots \rho_{\text{N-1}});$$
$$H_{l1} = (\rho_0, \rho_1, \rho_2 \ldots \rho_{\text{N-1-}l}), \qquad H_{l2} = (\rho_l, \rho_{l+1}, \rho_{l+2} \ldots \rho_{\text{N-1}}).$$

The autocorrelation coefficients $C16$ and $C_{\text{H}}(l)$ are defined:

$$C16 = cor\,(\text{H}_e, \text{H}_\text{o}) \qquad (2\text{-}6)$$

$$C_{\text{H}}(l) = cor\,(\text{H}_{l1}, \text{H}_{l2}) \qquad (2\text{-}7)$$

Set $l$ = 1, 2, 3 and 4, the features $C17$ to $C20$ are:

$$C17 = C_{\text{H}}(1), \quad C18 = C_{\text{H}}(2), \quad C19 = C_{\text{H}}(3), \quad C20 = C_{\text{H}}(4).$$

Besides the features mentioned above, we consider the difference between the testing image and the denoised. CI denotes the cover image and SI denotes the stego-image. Embedding information into images may be modeled as the process of adding noise. D (·)

is some denoising function. We define the difference between pre-denoised and post-denoised as follows:

$$E_{CI} = CI - D(CI) \tag{2-8}$$

$$E_{SI} = SI - D(SI) \tag{2-9}$$

Generally, the statistics of $E_{CI}$ and $E_{SI}$ are different. We apply wavelet hard-threshold denoising without shrinkage [Mallat, 1999] to the image. Firstly, apply wavelet transform to test image, set zero to the wavelet coefficients in HL, LH and HH sub-bands of which the absolute value are smaller than some threshold value *t,* and reconstruct the image by applying the inverse wavelet transform to the modified wavelet coefficients. The difference between the original and the reconstructed $E_t$ is the $m \times n$ matrix, $E_t = \{e_t^{ij}\}(i = 1, 2, ..., m; j = 1, 2, ..., n)$ . The correlation features in the difference domain are given as follows

$$C_E(t;\ k, l)\ =\ cor\ (E_{t,11(m-k)(n-l)},\ E_{t,(k+1)(l+1)mn}) \tag{2-10}$$

where,

$E_{t,11(m-k)(n-l)} = \{e_t^{ij}\}(i = 1, 2, ..., m-k; j = 1, 2, ..., n-l)$;

$E_{t,(k+1)(l+1)mn} = \{e_t^{ij}\}(i = k+1, k+2, ..., m; j = l+1, l+2, ..., n)$ .

Set different values to *t, k* and *l*, features *C*21 to *C*41 are presented as follows:

$C21 = C_E(1.5;\ 0,1)$; $C22 = C_E(1.5;\ 1,0)$; $C23 = C_E(1.5;\ 1,1)$; $C24 = C_E(1.5;\ 0,2)$;

$C25 = C_E(1.5;\ 2,0)$; $C26 = C_E(1.5;\ 1,2)$; $C27 = C_E(1.5;\ 2,1)$; $C28 = C_E(2;\ 0,1)$;

$C29 = C_E(2;\ 1,0)$; $C30 = C_E(2;\ 1,1)$; $C31 = C_E(2;\ 0,2)$; $C32 = C_E(2;\ 2,0)$;

$C33 = C_E(2;\ 1,2)$; $C34 = C_E(2;\ 2,1)$; $C35 = C_E(2.5;\ 0,1)$; $\mathrm{C}36 = C_E(2.5;\ 1,0)$;

C37=$C_E(2.5;\ 1,1)$; C38=$C_E(2.5;\ 0,2)$; C39=$C_E(2.5;\ 2,0)$; C40=$C_E(2.5;\ 1,2)$;

C41=$C_E(2.5;\ 2,1)$.

In RGB color images, the matrices $M_{\mathrm{r1}}$, $M_{\mathrm{g1}}$, and $M_{\mathrm{b1}}$ stand for the least significant bit planes of red, blue and green channels, respectively, the correlation coefficients $C_{\mathrm{rg}}$, $C_{\mathrm{rb,}}$ and $C_{\mathrm{gb}}$ are given as follows, where *abs*(·) denotes the absolute value function.

$$C_{\mathrm{rg}} = abs(cor(M_{\mathrm{r1}}, M_{\mathrm{g1}})) \quad (2\text{-}11)$$

$$C_{\mathrm{rb}} = abs(cor(M_{\mathrm{r1}}, M_{\mathrm{b1}})) \quad (2\text{-}12)$$

$$C_{\mathrm{gb}} = abs(cor(M_{\mathrm{g1}}, M_{\mathrm{b1}})) \quad (2\text{-}13)$$

Similar to (2-10), $E_{t,c}$ (*c=r, g, b*) is the difference across the color channels (red, green, and blue) of the original and the reconstructed. The correlation features are defined as follows.

$$C_{E_{rg}}(t) = cor(E_{t,r}, E_{t,g}) \quad (2\text{-}14)$$

$$C_{E_{rb}}(t) = cor(E_{t,r}, E_{t,b}) \quad (2\text{-}15)$$

$$C_{E_{gb}}(t) = cor(E_{t,g}, E_{t,b}) \quad (2\text{-}16)$$

After extracting the features defined above, we apply analysis of variance (ANOVA) [Avcibas *et al*, 2003; Rencher, 2002] to the features and pick up the features with high statistical significances as the final detectors.

## 2.5 One-way ANOVA

The purpose of one-way ANOVA [Rencher, 2002] is to determine whether the groups are actually different in the measured characteristic. The model of one-way ANOVA is:

$$y_{ij} = \overline{Y}_j + \varepsilon_{ij}, \quad (i = 1, 2, ..., I;\ j = 1, 2, ..., J) \tag{2-17}$$

$$\overline{Y}_j = \frac{1}{I}\sum_{i=1}^{I} y_{ij} \tag{2-18}$$

$$\overline{Y} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \tag{2-19}$$

Where $y_{ij}$ is a matrix of observations in which each column represents a different group and $\varepsilon_{ij}$ is a matrix of random disturbances. I is the sample number for every group and J is the number of groups. The variations, SS(Between) and SS(Within), are measured by:

$$SS(Between) = \sum_{j=1}^{J} (\overline{Y}_j - \overline{Y})^2 \tag{2-20}$$

$$SS(Within) = \sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \overline{Y})^2 \tag{2-21}$$

Dividing the corresponding sum of squares by its degrees of freedom, the mean sum of squares is given by:

$$MS(Between) = \frac{SS(Between)}{J - 1} \tag{2-22}$$

$$MS(Within) = \frac{SS(Within)}{IJ - J} \tag{2-23}$$

The F-statistic for ANOVA is a ratio of MS(Between) to MS(Within). The p-value is given by comparing the F-statistic with the F(J-1, IJ-J)-distribution, which tells the probability $H_{0:}$

$$\widetilde{F} = \frac{MS(Between)}{MS(Within)} \tag{2-24}$$

$$p = \mathrm{P}(F > \widetilde{F}) \qquad (2\text{-}25)$$

$$H_0 : \overline{Y_1} = \overline{Y_2} = \ldots = \overline{Y_J} \qquad (2\text{-}26)$$

## 2.6 Experimental Setup

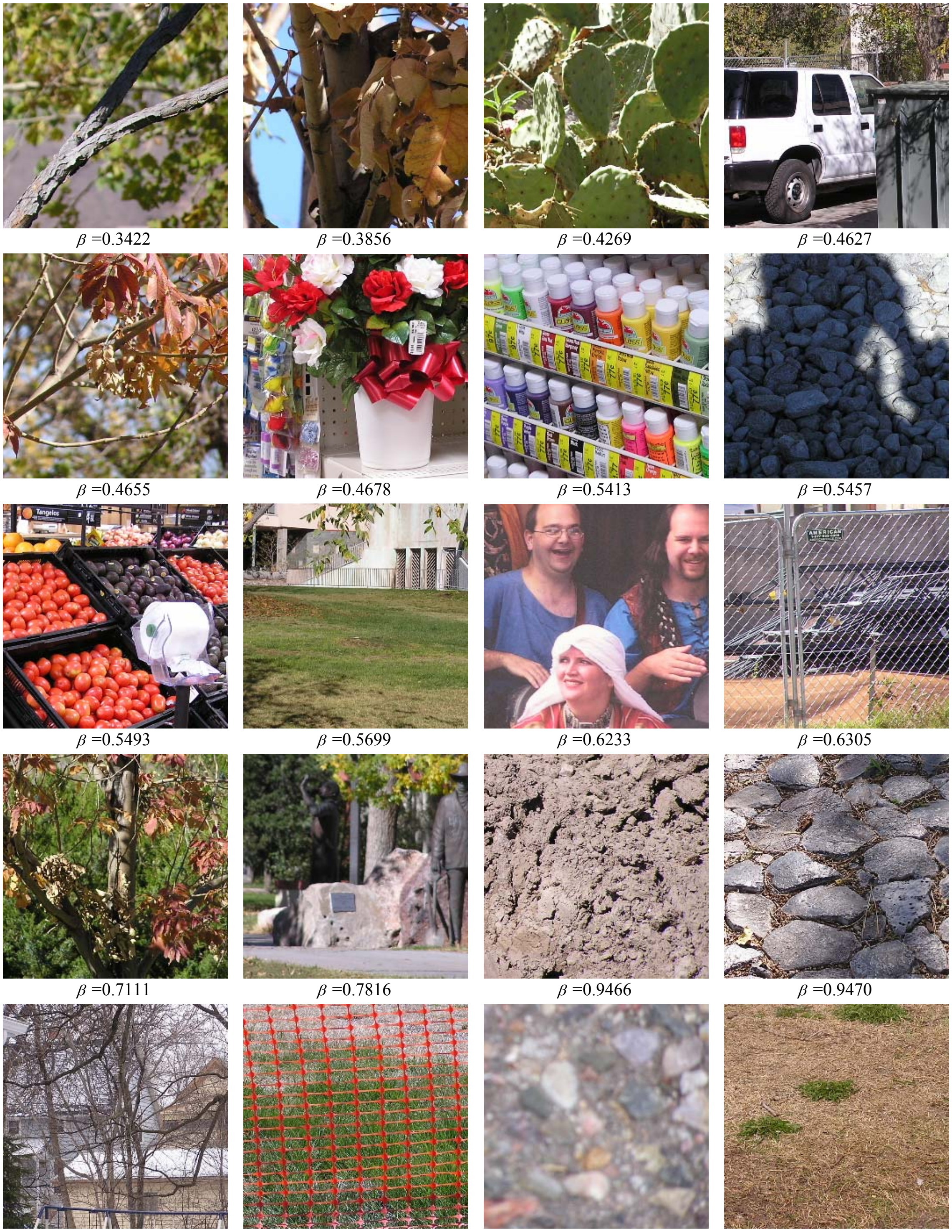
Generally, embedding data in once compressed images by modifying the pixel values is easier to detect than hiding data in uncompressed images. The original covers in our experiments are 5000 TIFF raw format digital pictures during 2003 to 2005. These images are 24-bit, 640×480 pixels, lossless true color and never compressed.

In steganalysis of color images, according to the pre-processing method in [Lyu and Farid, 2004, 2005], we cropped the original images into 256×256 pixels in order to get rid of the low complexity parts of the images. The cropped images are covers in our experiments. We categorize the covers according to the parameter of image complexity. The image complexity for color images is calculated as follows:

$$\beta = (\beta_r + \beta_g + \beta_b)/3 \qquad (2\text{-}27)$$

The variable $\beta_c (c = r, g, b)$ is the shape parameter of the GGD of the HH sub-band coefficients, corresponding to red, green, and blue channel. Fig. 2-3 lists some cover samples with different image complexity in color images.

In steganalysis of grayscale images, the cropped color images are converted into grayscales that are used as covers. The image complexity for grayscale is measured by the shape parameter of the GGD of the HH sub-band coefficients.



$\beta$ =0.3422 $\beta$ =0.3856 $\beta$ =0.4269 $\beta$ =0.4627

$\beta$ =0.4655 $\beta$ =0.4678 $\beta$ =0.5413 $\beta$ =0.5457

$\beta$ =0.5493 $\beta$ =0.5699 $\beta$ =0.6233 $\beta$ =0.6305

$\beta$ =0.7111 $\beta$ =0.7816 $\beta$ =0.9466 $\beta$ =0.9470

$\beta$ =1.0013 $\beta$ =1.2104 $\beta$ =1.5276 $\beta$ =1.6008

Fig. 2-3 Cover samples with different image complexity that is measured by the GGD shape parameter in the wavelet domain.

Stego-images are produced with the use of LSB matching algorithm. The hidden messages cover different types such as digital image, audio, text file, pdf file, zipped file, executable code, source code, random signal, etc. The hidden data in any two covers are different.

In steganalysis of color images, the correlation feature set consists of the following features: $C1$, $C2$, $C6$, $C10$, $C14$, $C15$, $C16$, $C17$, $C_E(2.5;\ 0,1)$, $C_E(2.5;\ 1,0)$, $C_E(2.5;\ 1,1)$, $C_E(3;\ 0,1)$, $C_E(3;\ 1,0)$, and $C_E(3;\ 1,1)$ defined in Section 3, corresponding to red, green, and blue channels, 14×3 = 42 features; $C_{E_{rg}}(t)$, $C_{E_{rb}}(t)$, $C_{E_{gb}}(t)$ ($t = 1$, 1.5, and 2), 3×3 = 9 features; in addition to $C_{\mathrm{rg}}$, $C_{\mathrm{rb}}$, and $C_{\mathrm{gb}}$, total 54 features. We compare the proposed feature set against other well-known feature sets of the Histogram Characteristic Function Center of Mass (HCFCOM) [Harmsen and Pearlman, 2003] and High-Order Moment statistics in Multi-Scale decomposition domain (HOMMS) [Lyu and Farid, 2004, 2005]. There are 3 features of HCFCOM and 216 features of HOMMS in color images.

The experiments on steganalysis of LSB matching steganography in grayscale images are the same to color images except that correlation feature set consists of the 41 features, C1 to C41, defined in section 3 and HOMMS feature set consists of 72 features in grayscale images. We extend HCFCOM feature set to the high order moments. Here HCFHOM stands for HCF center of mass High Order Moments; HCFHOM ($r$) denotes the $r^{\mathrm{th}}$ order moment. In our experiments, the HCFHOM feature set consists of HCFCOM and

HCFHOM($r$) ($r$ = 2, 3, and 4). Additionally, Ker proposed two novel ways of applying the HCF: calibrating the output using a down-sampled image and computing the adjacency histogram instead of the usual histogram [Ker, 2005]. The best discriminators are Adjacency HCFCOM (A.HCFCOM) and Calibrated Adjacency HCFCOM (C.A.HCFCOM).

Generally, different classifiers have different classification performances on different feature sets. Considering this point, we utilize the following classifiers:

1. Fisher Linear Discriminate (FLD),
2. Optimization of the Parzen Classifier (ParzenC),
3. Naive Bayes classifier (NBC),
4. Support Vector Machines (SVM),
5. Linear Bayes Normal Classifier (LDC),
6. Quadratic Bayes Normal Classifier (QDC),
7. Bayes Classifier (BC) that is based on maximal likelihood estimation of Gaussian mixture model,
8. Adaboost algorithm (Adaboost) which produces a classifier composed from a set of weak rules.

The details of these classifiers are described in the references [Duda, Hart and Stork, 2001; Friedman, Hastie and Tibshirani, 2000; Heijden *et al.*, 2004; Schlesinger and Hlavac, 2002; Taylor and Cristianini, 2004; Vapnik, 1998; Webb, 2002]. We apply each classifier to each feature set in each category of image complexity sixteen times. In each time, the training samples are randomly chosen and the remaining samples are tested.

## 2.7 Comparison of Statistical Significances

A parametric test is a test that requires a parametric assumption such as normality. Nonparametric test does not rely on parametric assumption like normality. Parametric tests work well with large samples even if the population is non-Gaussian [Motulsky, 1995]. Fig. 2-4 lists the F statistics and p-values of correlation features (CF), HOMMS, and HCFCOM features, extracted from 5000 covers and 5000 LSB matching stego-images in color images. The LSBP hiding ratio of these stego-images is 1 or the maximum hiding ratio. Fig. 2-4 indicates that, regarding individual features, HCFCOM features with the highest F statistics and lowest p-values are better than correlation features; correlation features with higher F statistics and lower p-vales are better than HOMMS features.  In HOMMS, there are many features with high p-values, indicating that these features are weak for discriminating cover images and stego-images. Regarding the F statistics of the correlation features, generally, the correlation features on inter-channels (feature-dimension 43 to 54) have higher F-statistics than the correlation features on intra-channels (feature-dimension 1 to 42), which exhibits that the correlation features on inter-channel are better than the intra-channel features.

Fig. 2-5 lists the F statistics and p-values of CF, HOMMS, HCFHOM, A. HCFCOM and C.A.HCFCOM features, extracted from 5000 covers and 5000 LSB matching stego-images in grayscale images of which the LSBP hiding ratio is 1 or the maximum hiding ratio. Regarding individual features, Fig. 2-5 indicates that correlation features with the highest F statistics and lowest p-values are better than other features; HOMMS features are not so good because the p-values of many HOMMS features are pretty high, indicating that the statistical significances of these HOMMS features are low and the classification performance is the worst.

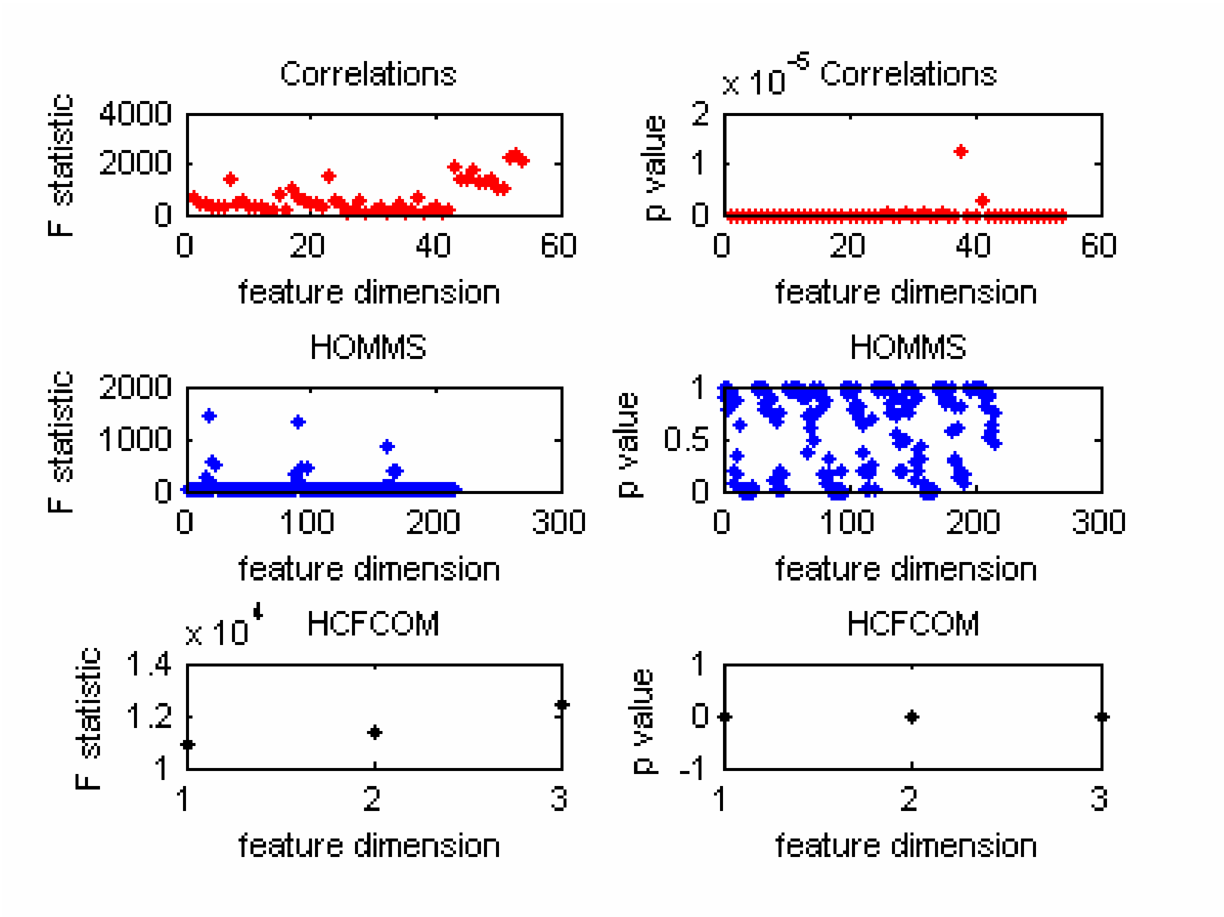


Fig. 2-4 F statistics and p-values of correlations, HOMMS, and HCFCOM features in color images.

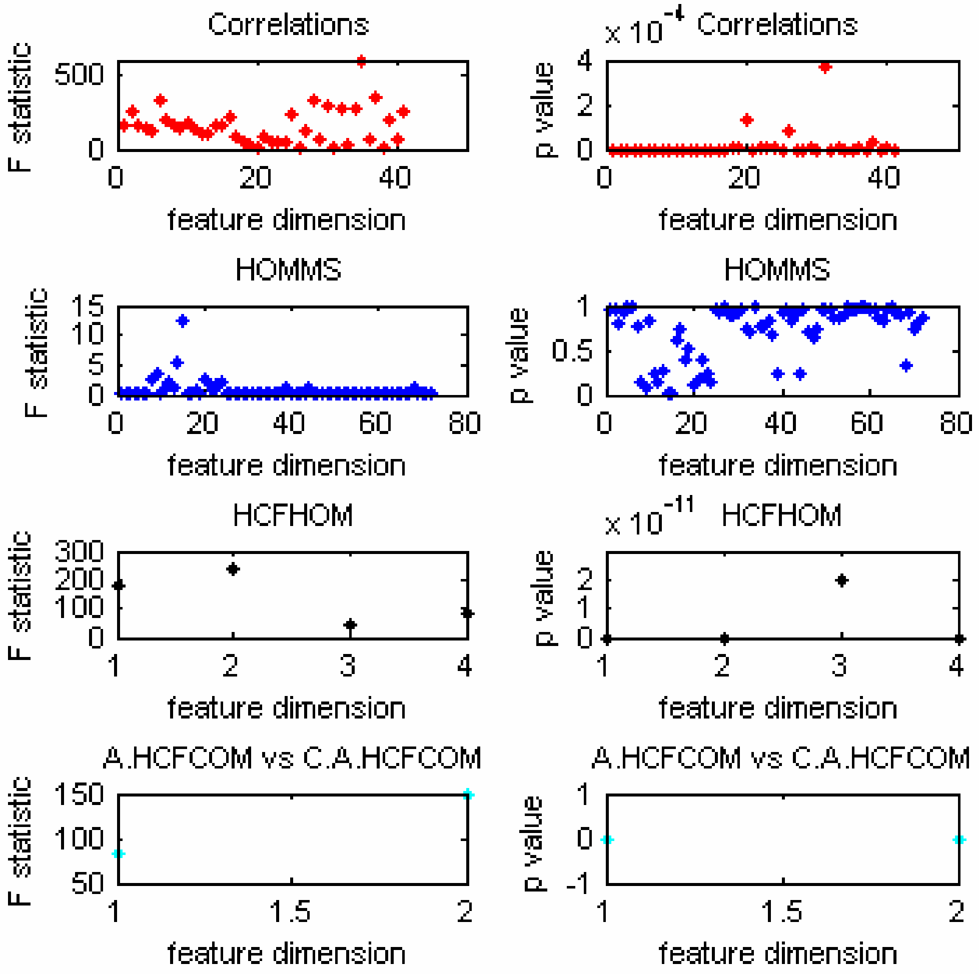


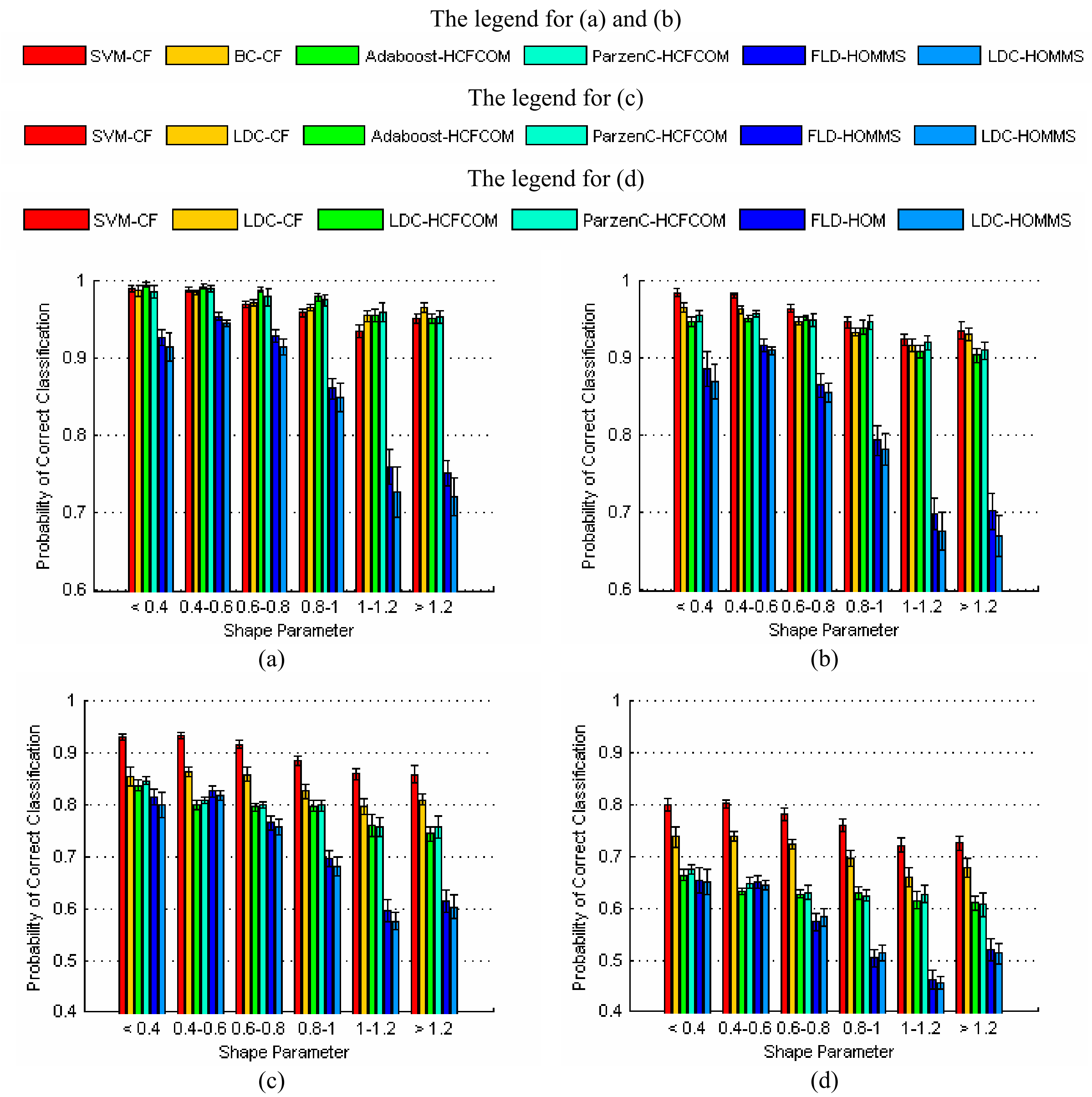
Fig. 2-5 F statistics and p-values of correlations, HOMMS, HCFHOM, A. HCFCOM and C.A.HCFCOM features in grayscale images.

## 2.8 Comparison of Classification Performances

Fig. 2-6 gives the top two classification accuracy (mean values and standard errors) on each feature set in color images under the LSBP hiding ratios of 1, 0.75, 0.5, and 0.25 (Fig. 8a-d)). Results show that, on the average, the set of correlation features (CF) outperforms HCFCOM and HOMMS; as the image complexity increases, the detection performances decrease; as the information-hiding ratio decreases, the diction performances decreases. Especially the detection performance of HOMMS decreases obviously and the classification performance is not good while the parameter of image complexity $\beta$ is bigger than one.

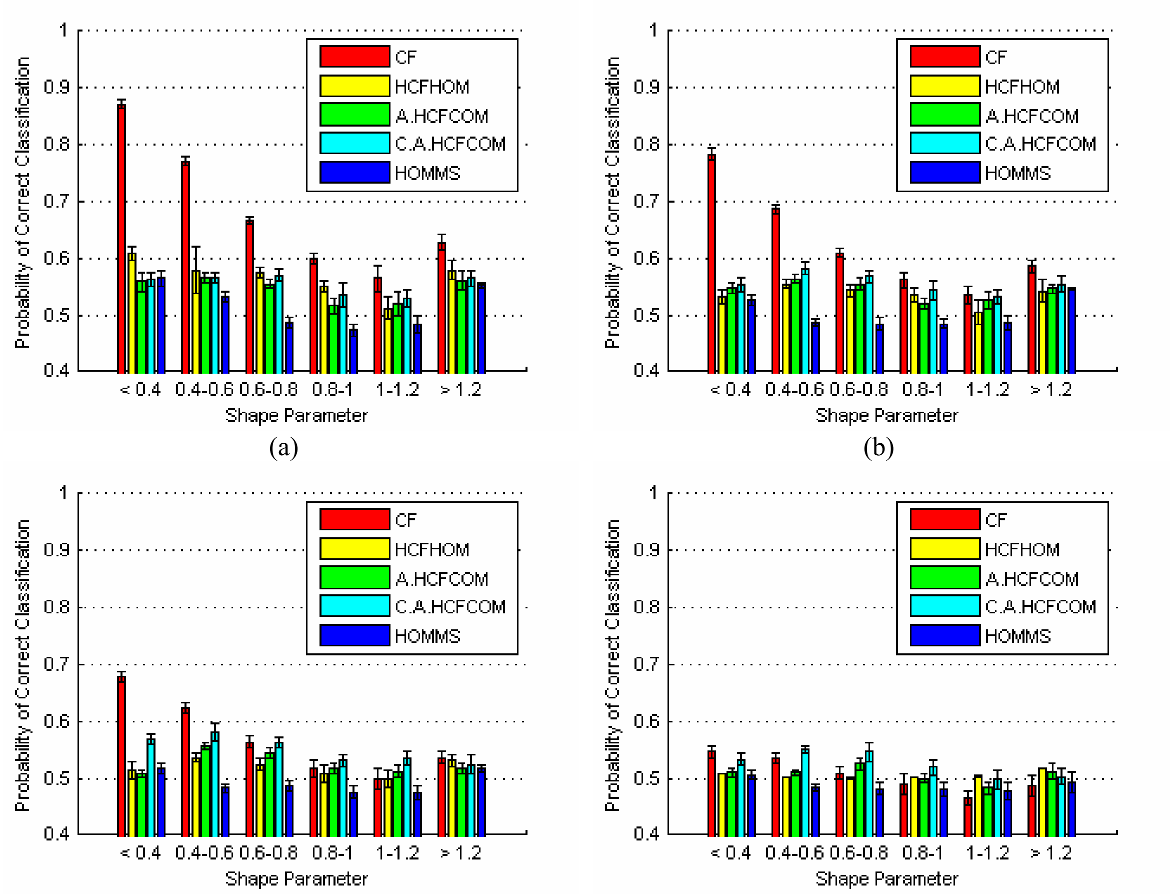
Fig. 2-7 lists the best classifications in grayscale images under the different hiding ratios and different image complexities. Results show that the detection performance of CF is the best and the performance of HOMMS is the worst, which is consistent with the analysis of the statistical significance. On the average, the classification performances decrease as the image complexity increases. When the parameter of image complexity is bigger than 0.8 or the LSBP hiding ratio is 0.25, the performances are not good. It obviously demonstrates that the steganalysis of LSB matching steganography in grayscale images is still very challenging in the cases where the grayscale image consists of complicated texture or the hiding ratio is very low.

In signal detection theory, a receiver operating characteristic (ROC) is a graphical plot of the sensitivity (fraction of true positives - TP) vs. 1-specificity (the fraction of false positives - FP) for a binary classifier system as its discrimination threshold is varied. The ROC curves under different image complexities in color images with the LSBP hiding ratios of 0.75 (I) and 0.5 (II) are shown in Fig. 2-8. Obviously, CF outperforms HCFCOM and HOMMS. The detection performances closely depend not only on the measure of information hiding ratio but also on the parameter of image complexity. As information hiding ratio decreases and image complexity increases, the detection performances decrease.



(a) (b) (c) (d)

Fig. 2-6 The best two classifications (mean values and standard errors) on each feature set (steganalysis of color LSB matching steganography). LSBP hiding ratios are 1(a), 0.75(b), 0.5(c), and 0.25(d), respectively. In the legends for (a), (b), (c), and (d), SVM-CF denotes applying SVM to Correlation Features (CF), Adaboost-HCFCOM denotes applying Adaboost to HCFCOM features, and so on.



(a)

(b)

(c)

(d)

Fig. 2-7 The best classification (mean values and standard deviations) on each feature set (steganalysis of grayscale LSB matching steganography). The LSBP hiding ratios are 1(a), 0.75(b), 0.5(c), and 0.25(d), respectively.

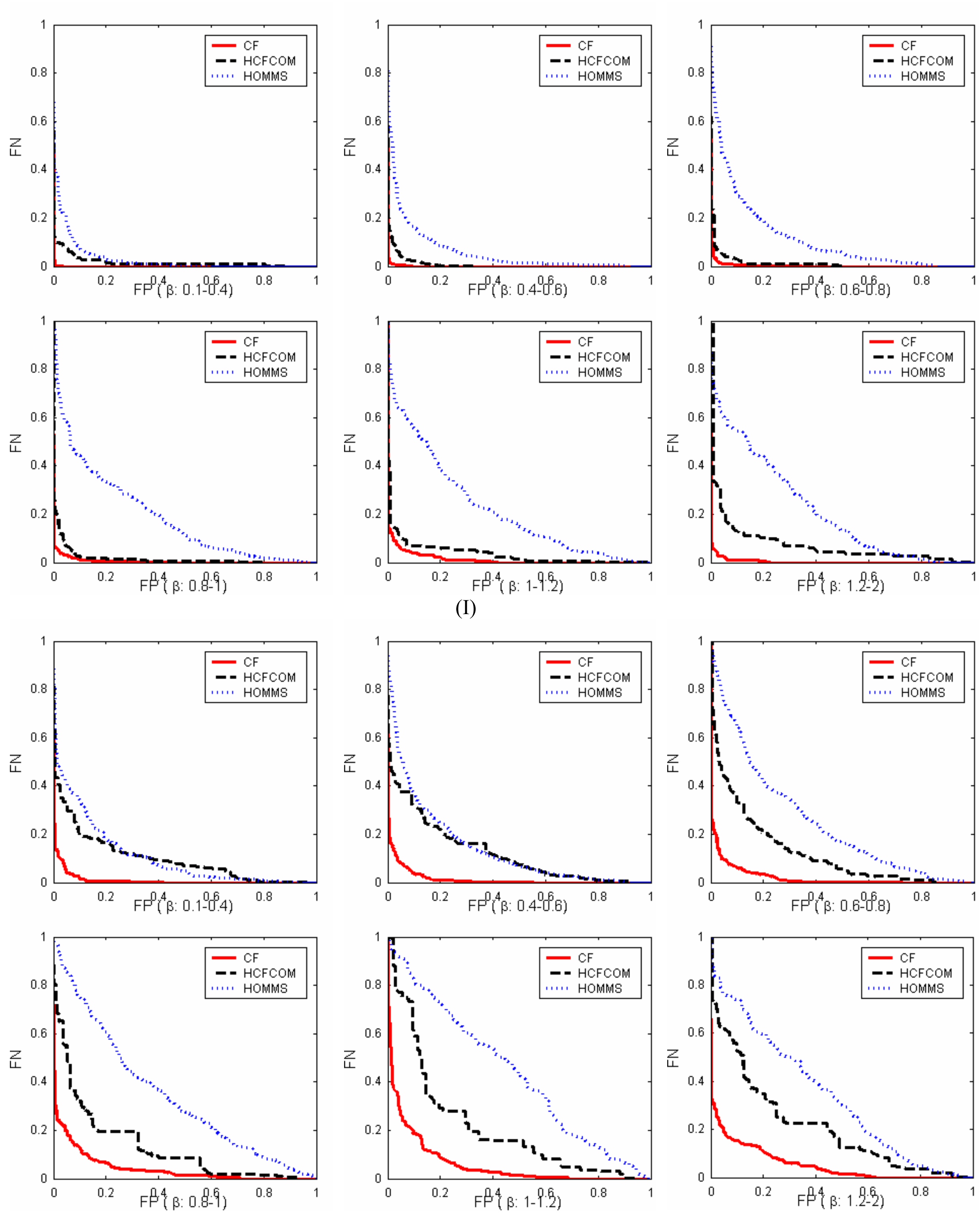
## 2.9 Discussions

All experiments show that the classification performances in color images are better than grayscale images. Fig. 2-4 reveals the statistical significances of the inter-channel correlation features are higher than intra-channel correlation features. In our point of view, on the average, there is stronger correlation in inter-channel than intra-channel which causes this result. Fig. 2-9(a) shows a color image and Fig. 2-9(b) presents the converted grayscale. Fig. 2-9(c), (e) and (g) are the joint probability of the red-green, red-blue and green-blue channels of the color image. Fig. 2-9(d), (f) and (h) are the joint probability of the adjacent pixels in the horizontal, vertical and diagonal directions of the grayscale image. The joint distribution of the grayscale is sparser, and the joint distribution of the color is more concentrated. The maximum values of the joint probability of the color are 0.012, 0.0030, and 0.0091, respectively, bigger than the maximum values of the grayscale.

As the image complexity increases, the variation of the adjacent pixels increases, and the correlation decreases. Fig. 2-10 shows two grayscale images with low parameter of image complexity (Fig. 2-10(a)) and high parameter (Fig. 2-10(b)). Fig. 2-10(c), (e), and (g) give the joint distribution of the adjacent pixels of Fig. 2-10(a); Fig. 2-10(d), (f), and (h) give the joint distribution of the adjacent pixels of Fig.2-10(b). The correlation information of the adjacent pixels of Fig.2-10(a) is stronger than Fig.2-10(b). It indicates that, with an increase of the parameter of image complexity, an increase of the variation of adjacent pixels results in decreasing both the detection performance and statistical significance.

Table 2-1 The ranksum test of the image complexity of the covers and the stego-images.

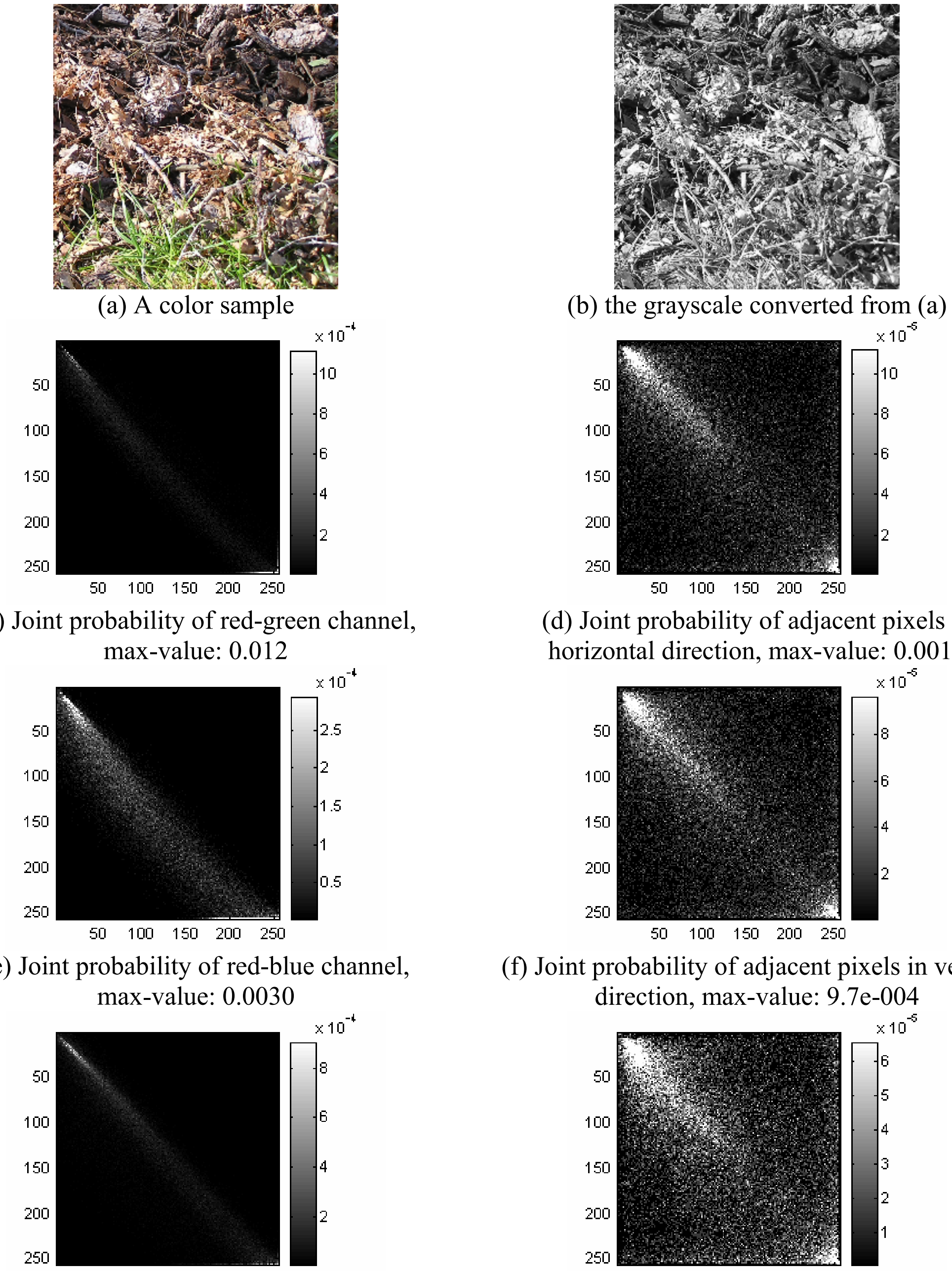
| Shape parameter $\beta$ | | < 0.4 | 0.4 ~ 0.6 | 0.6 ~ 0.8 | 0.8 ~ 1.0 | 1.0 ~ 1.2 | > 1.2 |
|---|---|---|---|---|---|---|---|
| Sample number | cover | 766 | 1576 | 982 | 770 | 515 | 391 |
| | stego | 636 | 1596 | 989 | 774 | 494 | 511 |
| Wilcoxon rank sum test | p-value | 0.0561 | 0.3319 | 0.5685 | 0.5273 | 0.8095 | 4.01e-008 |
| | HP | 0 | 0 | 0 | 0 | 0 | 1 |



(I)

(II)

Fig. 2-8 ROC curves in the steganalysis of LSB matching steganography in color images at the LSBP hiding ratios of 0.75 (I) and 0.5 (II). X-label gives the False Positive (FP) and y-label gives the False Negative (FN). The shape parameter β at the bottom of each figure indicates the range of the image complexity under the experiment.



(a) A color sample

(b) the grayscale converted from (a)

(c) Joint probability of red-green channel, max-value: 0.012

(d) Joint probability of adjacent pixels in horizontal direction, max-value: 0.0011

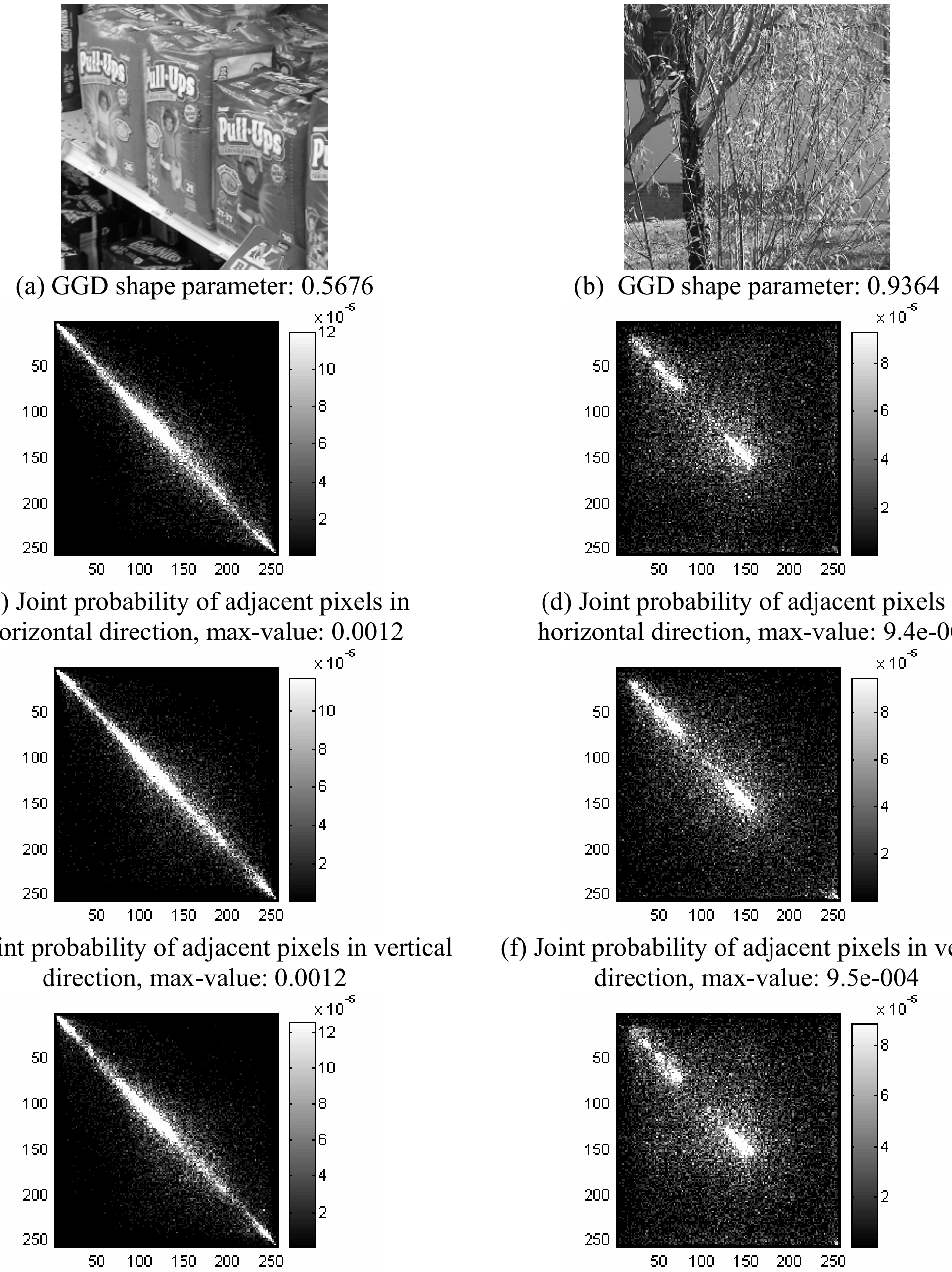
(e) Joint probability of red-blue channel, max-value: 0.0030

(f) Joint probability of adjacent pixels in vertical direction, max-value: 9.7e-004

(g) Joint probability of blue-green channel, max-value: 0.0091

(h) Joint probability of adjacent pixels in diagonal direction, max-value: 6.7e-004

Fig. 2-9 Comparison of correlation in color and grayscale images. Left column is a color sample and the correlations of the inter-channels; right column is the grayscale sample converted from (a) and the correlation of the adjacent pixels. It indicates that the correlation information on inter-channel is higher than that on intra-channel by comparing the joint probabilities in left column and the joint probabilities in right column.



(a) GGD shape parameter: 0.5676

(b) GGD shape parameter: 0.9364

(c) Joint probability of adjacent pixels in horizontal direction, max-value: 0.0012

(d) Joint probability of adjacent pixels in horizontal direction, max-value: 9.4e-004

(e) Joint probability of adjacent pixels in vertical direction, max-value: 0.0012

(f) Joint probability of adjacent pixels in vertical direction, max-value: 9.5e-004

(g) Joint probability of adjacent pixels in diagonal direction, max-value: 0.0013

(h) Joint probability of adjacent pixels in diagonal direction, max-value: 8.9e-004

Fig. 2-10 Comparison of correlations of low complexity and high complexity grayscales. Left column is a grayscale sample with low complexity and the correlations of the adjacent pixels; right column is the grayscale sample with high complexity and the correlation of the adjacent pixels. It indicates that the correlation information of the image with low complexity is higher than that of the image with high complexity.

Figures 2-6 and 2-7 show when shape parameter is bigger than 1.2, the classification suddenly improves (not so much). This seems to contradict with the conclusion that with increasing complexity detection ability decreases. Actually, this contradiction is caused by the different distribution of the image complexity of the covers and the stego-images in the case where the shape parameter is bigger than 1.2; it doesn't contract with the conclusion. To further explain this contradiction, we study the affection on the image complexity caused by information-hiding. A non-parametric test, Wilcoxon rank sum test is performed for equal medians at the 0.05 significance level. The hypothesis is that two independent samples X and Y (X and Y can be different lengths) come from distributions with equal medians, and returns the p-value, the probability of observing the given result, or one more extreme, by chance if the null hypothesis ("medians are equal") is true. Table 2-1 lists the sample-numbers of covers and steganograms in the grayscale image, the LSBP hiding ratio is 1. HP=0 indicates that the null hypothesis ("medians are equal") cannot be rejected at the 5% level. HP=1 indicates that the null hypothesis can be rejected at the 5% level. Generally, the LSB matching information-hiding will increase the image complexity, but not so much. In table 2-1, there are 766 cover samples and 636 stego-samples in the category of $\beta < 0.4$. It means that with information-hiding the image complexity increases, there are 130 stego- samples and the image complexity is bigger than 0.4, although the original image complexity is smaller than 0.4. Similarly, 120 images shift to the category of $\beta > 1.2$ from the category of $\beta < 1.2$ with the information-hiding. Wilcoxon rank sum test indicates that the shifting of the image complexity with the information-hiding don't change the distribution in the categories of $\beta < 1.2$, but the distribution is changed in the category of $\beta > 1.2$. Fig. 2-11 shows the distribution difference of the covers (on the left) and the stego-images (on the right) in the categories of β > 1.2 (on the upper) and 1 < β < 1.2 (on the lower) in the grayscale imaged, the LSBP hiding ratio is 1. Table 2-1 and Fig. 2-11 show that, it is the distribution difference that results in the contraction in Figures 2-6 and 2-7. Again, it indicates that the detection performance depends on the image complexity. Fig. 2-12 is the boxplot of the image complexity of the covers and the stego-images (grayscale), the LSBP hiding ratio is 1. It shows that the information-hiding will increase the image complexity, but the increase is small. In our experiments, when the shape parameter is smaller than 1.2, the information-hiding didn't change the distribution of the image complexity, but it did change the distribution of the image complexity in the case where the shape parameter is bigger than 1.2, which results in the contradiction in the results shown in figures 2-6 and 2-7.

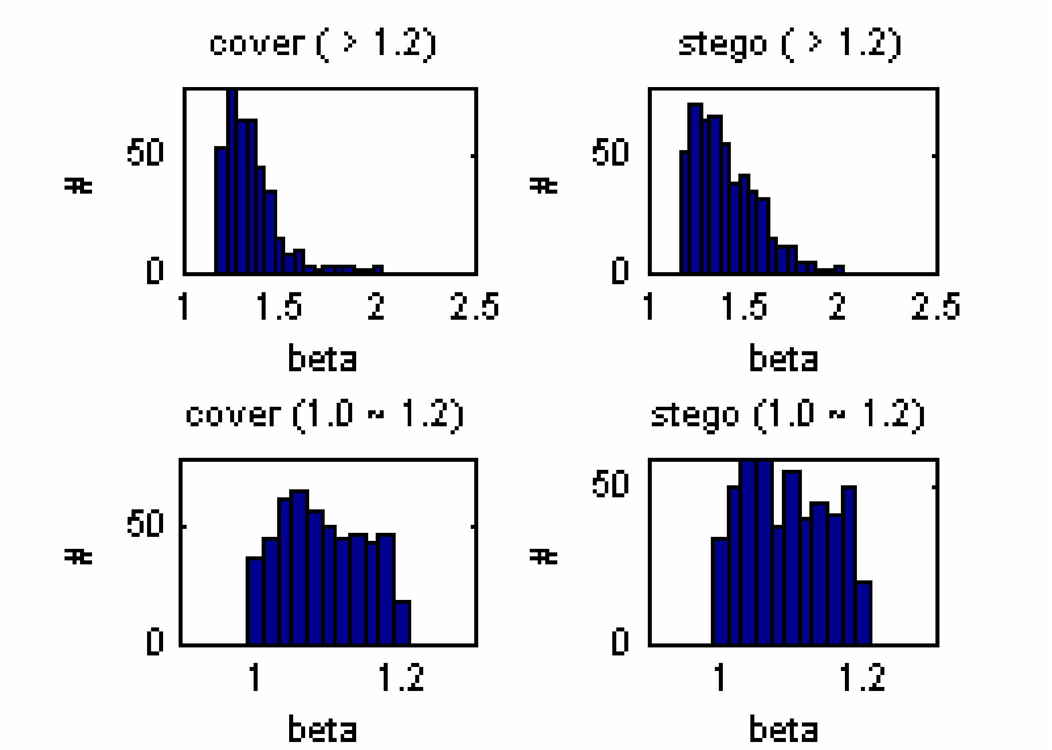


Fig. 2-11 The distribution of the image complexity of the covers and the stego-images (grayscale) in the categories of β > 1.2 and β in [1, 1.2].The LSBP hiding ratio is 1.

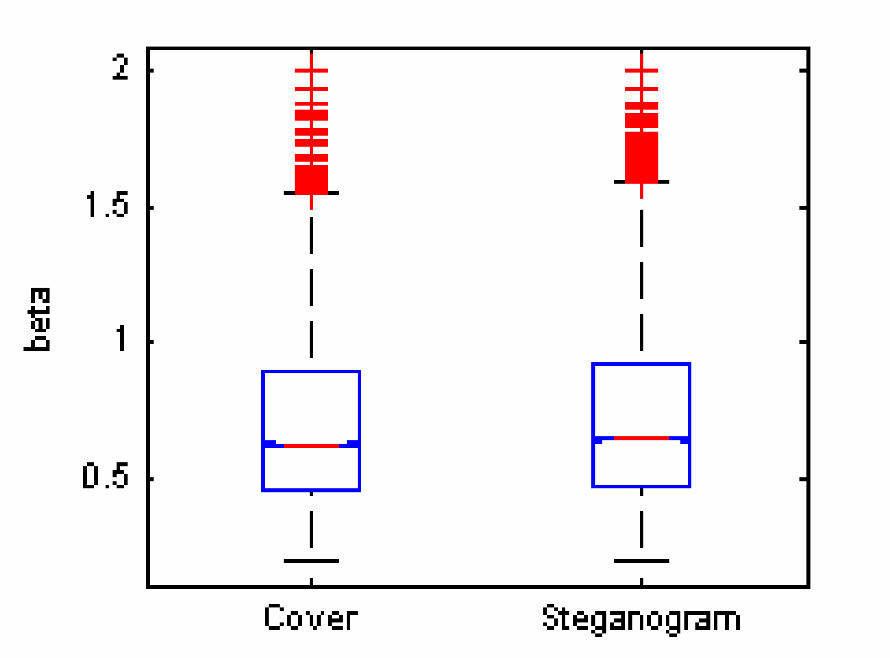


Fig. 2-12 The boxplot of the image complexity of the covers and the stego-images (grayscale). The LSBP hiding ratio is 1.

## 2.10 Conclusions

Information-hiding ratio is a well-known reference to evaluation of the performance of steganalysis. However, few publications clearly demonstrate the relation of image complexity and detection performance. In this chapter, we introduce the parameter of image complexity to the field of steganalysis and utilize the shape parameter of Generalized Gaussian Distribution (GGD) in the wavelet domain to measure the image complexity. To detect the presence of hidden data in LSB matching steganography, we present different correlation features. Comparing to other well-known features of HCFCOM and HOMMS in color images, and HCFHOM, HOMMS, A.HCFCOM, and C.A.HCFCOM in grayscale images, overall, our feature set performs the best. Experimental results show that the statistical significance of features and the detection performance closely depend not only on the information-hiding ratio but also on the image complexity. While the hiding ratio decreases and the image complexity increases, the significance and detection performance decrease. Meanwhile, the steganalysis of LSB matching steganography in grayscale images is still very challenging in the cases of complicated textures or low hiding ratios.

There is high correlation of adjacent pixels. Based on the features presented in this chapter, we also successfully applied the method to detecting the information-hiding behaviors in other space-hiding steganogrphic systems [Liu, Sung and Ribeiro, 2005; Liu, Sung and Xu, 2005] and the experimental results also support the hypothesis that the information-hiding in space-hiding steganographic systems affect the high correlation.

Feature selection is a general problem. This chapter did not cope with the issue of optimizing the feature set, which will be studied in the next chapter as well as the improvement of the detection of LSB matching steganography in grayscale images.

# CHAPTER 3: IMPROVED DETECTION OF LSB MATCHING IN GRAYSCALE IMAGES

## 3.1 Introduction

Many detection methods in steganalysis are based on feature mining and pattern classification techniques. Regarding feature mining, besides feature extraction, another general problem is feature selection. Analysis of variance (ANOVA) is utilized to choose good image quality metrics [Avcibas *et al*., 2003]. In detail, the higher the F statistic, the lower the p value, and the better the feature is. This feature selection is simple and runs fast. It is good in evaluating the statistical significance of the individual feature, but it doesn't consider the interaction of the features, and probably, the final feature set is not optimal. Otherwise there has been little research that deals with the feature selection problem with specific respect to steganalysis.

We introduced the shape parameter of Generalized Gaussian Distribution (GGD) in the wavelet domain to measure the image complexity and evaluate the steganalysis performance [Liu, Sung, Xu, Ribeiro, 2006]; although the method proposed therein is successful in detecting LSB matching steganography in color images and outperforms other well-known methods, its performance is not so good in grayscale images, which is generally more difficult and shown in chapter 2.

To improve the performance in detecting LSB matching steganography in grayscale images, based on our previous work [Liu and Sung, 2007], in addition to correlation features described in the previous chapter, four new types of features are designed and a Dynamic Evolving Neural Fuzzy Inference System (DENFIS) [Kasabov and Song, 2002; Kasabov, 2002] is introduced in this chapter. We also adopt the feature selection of Support Vector Machine Recursive Feature Elimination (SVMRFE) [Guyon *et al.*, 2002; Liu and Sung, 2007] to choose the features in our steganalysis.

Comparing against other well-known methods in terms of steganalysis performance, the new feature set performs the best. DENFIS is superior to other compared learning classifiers including SVM and adaboost. SVMRFE outperforms DENFIS based sequential forward selection and statistical significance based feature selection like T-test.

Our experimental results also indicate that image complexity is an important parameter to evaluation of the detection performance. At a certain information-hiding ratio, it is much more difficult to detect the information-hiding behavior in high image complexity than that in low complexity.

## 3.2 Feature Extraction

### 3.2.1 Entropy and High Order Statistics of the Histogram of the Nearest Neighbors

As shown in Fig. 2-2, there is high correlation of the adjacent pixels in ordinary images and we have a hypothesis that the information-hiding behavior will affect the joint distribution of the adjacent pixels. Based on this hypothesis, we consider the statistics of the histogram of the nearest neighbors. In chapter 2, Fig. 2-2 just shows a 2-D case of the nearest neighbors. Here we consider a 3-D case. The grayscale value at the point ($i$, $j$) is represented by $x$, the grayscale value at the point ($i$+1, $j$) is $y$, and the grayscale value at the point ($i$, $j$+1) is $z$. The variable $\boldsymbol{H}(x, y, z)$ denotes the occurrence of the pair ($x, y, z$) of the image, or the histogram of the nearest neighbors (NNH).

The entropy of NNH (NNH_E) is calculated as follows:

$$NNH\_E = -\sum \rho_H \log_2 \rho_H \tag{3-1}$$

Where $\rho_H$ denotes the distribution density of the NNH. The symbol $\sigma_H$ denotes the standard deviation of $\boldsymbol{H}$ (or NNH). The $r^{th}$ high order statistics of NNH (NNH_HOS) is given as:

$$NNH\_HOS(r) = \frac{\frac{1}{N^3}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\sum_{z=0}^{N-1}\left(H(x,y,z)-\frac{1}{N^3}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\sum_{z=0}^{N-1}H(x,y,z)\right)^r}{\sigma_H{}^r} \tag{3-2}$$

Where N is the number of possible gray scales of the image, e.g., for an 8-bit grayscale image, N = 256.

### 3.2.2 Probabilities of the Equal Neighbors

Besides the features on the histogram of the nearest neighbors, the probabilities of the equal neighbors are extracted. The structures of the equal neighbors are shown in Fig. 3-1, where $a$ represents the pixel value. Equal neighbors mean that the pixel values in the structure equal to each other.

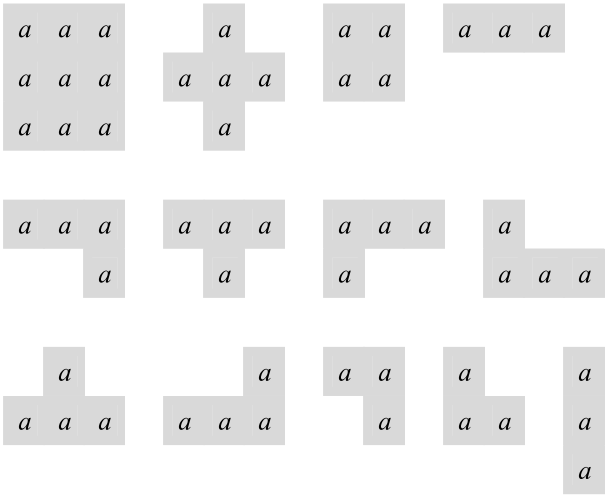


Fig. 3-1 The structures of the equal neighbors.

### 3.3 Introduction to DENFIS

Neuron-fuzzy inference systems consist of a set of rules and an inference method that are embodied or combined with a connectionist structure for better adaptation. Evolving neuron-fuzzy inference systems are such systems, where both the knowledge and the mechanism evolve and change in time, with more examples presented to the system [Kasabov 2002]. The dynamic evolving neuron-fuzzy inference system, or DENFIS [Kasabov and Song, 2002], uses the Takagi-Sugeno type of fuzzy inference method [Takagi and Sugeno, 1985]. The inference used in DENFIS is performed on *m* fuzzy rules indicated as follows:

$$\begin{aligned}
&\textit{If } x_1 \textit{ is } R_{11} \textit{ and } x_2 \textit{ is } R_{12} \textit{ and } \ldots \textit{ and } x_q \textit{ is } R_{1q}, \textit{ then } y \textit{ is } f_1(x_1, x_2, \ldots, x_q) \\
&\textit{If } x_1 \textit{ is } R_{21} \textit{ and } x_2 \textit{ is } R_{22} \textit{ and } \ldots \textit{ and } x_q \textit{ is } R_{2q}, \textit{ then } y \textit{ is } f_2(x_1, x_2, \ldots, x_q) \\
&\ldots \ldots \\
&\textit{If } x_1 \textit{ is } R_{m1} \textit{ and } x_2 \textit{ is } R_{m2} \textit{ and } \ldots \textit{ and } x_q \textit{ is } R_{mq}, \textit{ then } y \textit{ is } f_m(x_1, x_2, \ldots, x_q)
\end{aligned}$$

Where "$x_j$ is $R_{ij}$", $i$ = 1,2, …, $m$; $j$ = 1,2,…, $q$, are $m \times q$ fuzzy propositions that form $m$ antecedents for $m$ fuzzy rules respectively; $x_j$, $j$ = 1, 2, … , $q$, are antecedent variables defined over universes of discourse $X_j$, $j$ = 1, 2, … , $q$, and $R_{ij}$, $i$ = 1, 2, …, $m$; $j$ = 1, 2, … , $q$ are fuzzy sets defined by their fuzzy membership functions: $X_j \rightarrow [0,1]$, $i$ = 1, 2, …, $m$; j = 1, 2, …, $q$. In the consequent parts of the fuzzy rules, $y$ is the consequent variable, and crisp functions $f_i$, $i$ = 1, 2, … , $m$, are employed.

In the DENFIS model, all fuzzy membership functions are triangular type functions defined by the three parameters, *a, b,* and *c,* as given below:

$$\mu(x) = mf(x,a,b,c) = max(min((x\text{-}a)/(b\text{-}a), (c\text{-}x)/(c\text{-}b)), 0) \qquad (3\text{-}3)$$

Where $b$ is the value of the cluster centre on the x dimension, $a = b - d \times Dthr$, $d$ = 1.2 ~ 2. The threshold value, *Dthr*, is a clustering parameter.

For an input vector $x^0 = [x_1^0 \ x_2^0 \cdots x_q^0]$, the result of the inference, $y^0$, or the output of the system, is the weighted average of each rule's output indicated as follows:

$$y^0 = \frac{\sum_{i=1}^{m} w_i f_i(x_1{}^0, x_2{}^0, ..., x_q{}^0)}{\sum_{i=1}^{m} w_i} \qquad (3\text{-}4)$$

where, $w_i = \prod_{j=1}^{q} R_{ij}(x_j{}^0); i = 1,2,...,m; j = 1,2,...,q.$

In the DENFIS on-line model, the first-order Takagi-Sugeno type fuzzy rules are employed. In the DENFIS off-line models, the first-order and an extended high-order Takagi-Sugeno inference engines are used, corresponding to a linear model and an MLP-based model, respectively. The experiments indicate that the DENFIS with MLP-based model has the best prediction performance. The details of the DENFIS off-line learning process is presented in the reference [Kasabov, 2002].

## 3.4 Feature Selection in Steganalysis

To detect the information-hiding behaviors in steganography, many articles proposed different features or measures. In steganalysis, feature selection should be a general problem; to our knowledge, however, few publications cope with this issue except Avcibas *et al.* presented a universal steganalysis based on image quality metrics and utilized analysis of variance (ANOVA) to choose the good measures [Avcibas *et al.*, 2003]. Essentially, this feature selection belongs to filtering approach and the final feature set may not be optimal.

Generally, feature selection can be grouped into three categories: filtering, wrapper methods and embedded methods. Filter methods select feature subsets independently from the learning classifiers and do not incorporate learning [Xu and Chen, 2005; Pvlidis and Noble, 2001]. A weakness of filtering methods is that they just consider the individual feature in isolation and ignore the possible interaction of features among them. Yet, the combination of these features may have a combined effect that does not necessarily follow from the individual performance of features in the group. If there is a limit on the number of features to be chosen, we may not be able to include all informative features.

Wrapper methods wrap around a particular learning algorithm that can assess the selected feature subsets in terms of the estimated classification errors and then build the final the final classifiers [Inza *et al*., 2002]. One of the well-known methods is Support Vector Machine Recursive Feature Elimination (SVMRFE), which refines the optimum feature set by using SVM in a wrapper approach to address the problem of gene selection in the analysis of microarray data [Guyon *et al*., 2002]. Additionally, Sequential Forward Selection (SFS) is a greedy search algorithm in wrapper methods. To deal with the issue of feature selection in our steganalysis, we compare these three feature selections: DENFIS based SFS (DENFIS-SFS), SVMRFE, and T-test, a filtering feature selection which is similar to the ANOVA approach in steganalysis [Avcibas *et al*., 2003].

## 3.5 Experiments and Results

### 3.5.1 Experimental Setup

The original images in our experiments are 5000 TIFF raw format digital pictures, taken in USA during 2003 to 2005. These images are 24-bit, 640×480 pixels, lossless true color and never compressed. As mentioned in chapter 2, we cropped the original images into 256×256 pixels in order to get rid of the low complexity parts of the images. The cropped color images are converted into grayscales and LSB matching stego-images are produced by hiding data in these grayscales. The hiding ratio (the ratio of the file size of the hidden data to the file size of the cover image) is 12.5%. The hidden data in any two images are different.

We categorize the grayscale images (covers and stego-images) according to the image complexity which is measured by the shape parameter $\beta$ of the GGD of the HH wavelet sub-band coefficients. Fig. 3-2 lists some cover samples with different shape parameters in our experiments.

### 3.5.2 Feature Extraction and Comparison

The following features are extracted:

1. Shape parameter $\beta$ of the GGD of the HH wavelet sub-band that measures the image complexity.
2. Entropy of the histogram of the nearest neighbors, NNH_E, defined in (3-1).
3. The high order statistics of the histogram of the nearest neighbors, NNH_HOS($r$) in (3-2), and $r$ is set from 3 to 22, total 20 high order statistics.
4. Probabilities of the equal neighbors (PEN), described in 3.2.2.
5. Correlations features defined in chapter 2: C1 in (2-3), C($k,l$) in (2-4), C2 in (2-6), $C_H(l)$ in (2-7), and $C_E(t; k,l)$ in (2-10).

   We set the following lag distance to $k$ and $l$ in C($k,l$) and get 14 features:

   1) $k = 0, l = 1, 2, 3,$ and $4; l = 0, k = 1, 2, 3,$ and 4.

   2) $k = 1, l = 1; k = 2, l = 2; k = 3, l = 3; k = 4$ and $l = 4$.

   3) $k = 1, l = 2; k = 2, l = 1$.

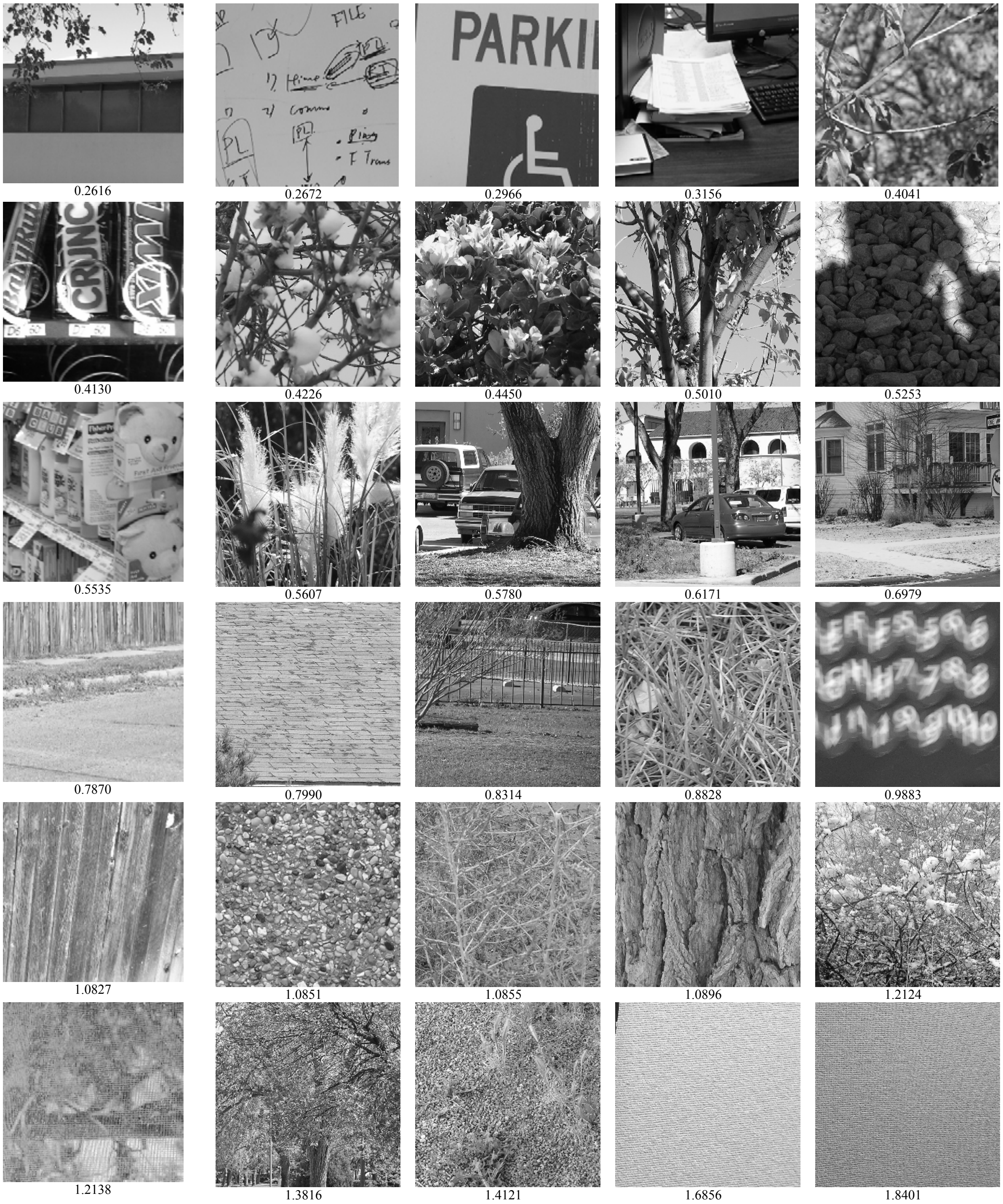


Fig. 3-2 Some cover samples (scaled) and the shape parameters.

In (2-7), $l$ is set to 1, 2, 3, and 4. In (2-10), we set the following lag distances to $k$ and $l$ in $C_E(t; k,l)$ and get following pairs:

$C_E(t; 0,1)$, $C_E(t; 0,2)$, $C_E(t;1,0)$, $C_E(t; 2,0)$, $C_E(t; 1,1)$, $C_E(t; 1,2)$, and $C_E(t; 2,1)$. $t$ is set 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5.

We record the fifth type of correlation features as CF; types 1 to 5 as EHPCC (Entropy, High order statistics, Probabilities of the equal neighbors, Correlation features, and Complexity).

To compare EHPCC with other well-known features, the Histogram Characteristic Function Center of Mass (HCFCOM) [Harmsen and Pearlman, 2003] is extracted because the hiding process of LSB matching steganography can be modeled in the context of additive noise. We extend the HCFCOM to the high order moments. HCFHOM stands for HCF center of mass High Order Moments; HCFHOM ($r$) denotes the $r^{th}$ order statistics. In our experiments, the HCFHOM feature set consists of HCFCOM and HCFHOM($r$) ($r$ = 2, 3, and 4). We also compare Adjacency HCFCOM (A.HCFCOM) and Calibrated Adjacency HCFCOM (C.A.HCFCOM) proposed by Ker [Ker, 2005]. Additionally, Farid and Lyu [Lyu and Farid, 2004, 2005] presented an approach to detecting hidden messages in images by building High-Order Moment statistics in Multi-Scale decomposition domain (we call these features HOMMS), which consists of 72-dimension features in grayscale images.

All the features mentioned above are listed in table 3-1.

Table 3-1 Proposed and compared features in our experiments.

| Feature set | Description of the features | The source | The number of Features |
|---|---|---|---|
| EHPCC | Entropy of NNH<br>( NNH_E ) | Defined in (3-1) | 1 |
| | High order statistics of NNH<br>( NNH-HOS (r), r = 3, 4, …, 22 ) | Defined in (3-2) | 20 |
| | Probabilities of Equal Neighbors<br>( PEN ) | Described in 3.2.2<br>Fig. 3-1 presents the structures of the equal neighbors. | 13 |
| | Correlation Features<br>( CF ) | C1 defined in (2-3);<br>C(k, l) defined in (2-4):<br>C(0, 1), C(0, 2), C(0, 3), C(0, 4), C(1,0), C(2, 0), C(3, 0), C(4,0), C(1,1), C(2, 2), C(3, 3), C(4, 4), C(1, 2), C(2, 1);<br>C2 defined in (2-6);<br>$C_H(l)$ in (2-7):<br>$C_H(1)$, $C_H(2)$, $C_H(3)$, $C_H(4)$;<br>$C_E(t; k,l)$ in (2-10):<br>$C_E(t; 0,1)$, $C_E(t; 0,2)$, $C_E(t;1,0)$, $C_E(t; 2,0)$, $C_E(t; 1,1)$, $C_E(t; 1,2)$, and $C_E(t; 2,1)$.<br>*t* is set 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. | 83 |
| | complexity measure | The shape parameter $\beta$ in (2-1)<br>[Sharifi and Leon-Garcia, 1995; Liu *et al.* 2006] | 1 |
| HCFHOM | HCFCOM<br>and the high order statistics<br>HCFHOM(r)<br>( r = 2, 3, 4) | [Harmsen and Pearlman, 2003] | 4 |
| A.HCFCOM | Adjacent HCFCOM | [Ker, 2005] | 1 |
| C.A.HCFCOM | Calibrated adjacent HCFCOM | [Ker, 2005] | 1 |
| HOMMS | High-order moment statistics in multi-scale decomposition domain | [Lyu and Farid, 2004 and 2005] | 72 |

Fig. 3-3 lists the F statistics and p-values of NNH_E and NNH_HOS, shape parameter $\beta$ and correlation features, Probabilities of the equal neighbors, HOMMS features, HCFHOM features, A. HCFCOM and C.A. HCFCOM features, extracted from the 5000 grayscale covers and the 5000 LSB matching stego-images. Fig. 3-3 indicates that, regarding the statistical significance, on the average, NNH-E, NNH-HOS, correlation features, and probabilities of the equal neighbors with high F statistics and very small p-values are better than HCFHOM, A. HCFCOM and C.A.HCFCOM features; and HOMMS features are not good because the p-values of most HOMMS features are high and the F statistics are small, it implies that the discrimination ability of HOMMS features is very weak. Fig. 3-3 also shows that the F statistic of the shape parameter $\beta$ is small and the p-value is close to 0, which means that the information-hiding changes the image complexity of the original cover, but the affection is very weak.

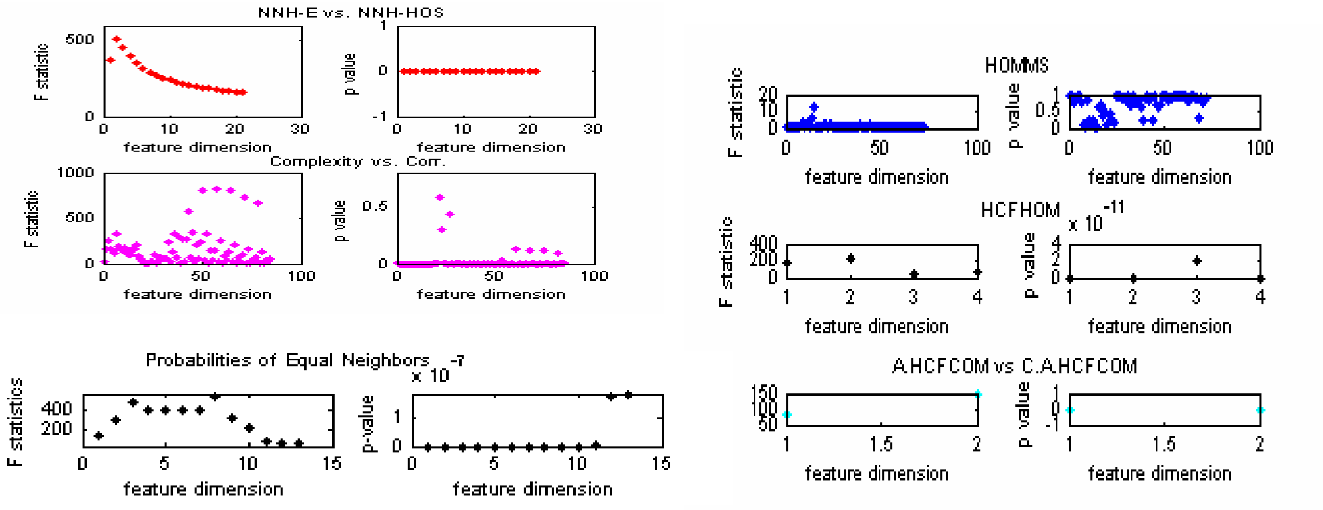


Fig. 3-3 F statistics and p-values of NNH-E (feature dimension 1 on the upper left), NNH-HOS (feature dimension 2 to 21 on the upper left), shape parameter β (feature dimension 1 on the middle left), correlation features (feature dimension 2 to 84 on the middle left), probabilities of equal neighbors, HOMMS, HCFHOM , A. HCFCOM, and C.A. HCFCOM features.

### 3.5.3 Detection Performance on Feature Sets

To compare the detection performances on these feature sets with different classifiers, besides DENFIS, we apply the following classifiers to each feature sets. These classifiers are Naive Bayes Classifier (NBC), Support Vector Machines (SVM), Quadratic Bayes Normal Classifier (QDC), and adaboost that produces a classifier composed from a set of weak rules [Friedman, Hastie and Tibshirani, 2000; Heijden et al., 2004; Vapnik, 1998; Schlesinger and Hlavac, 2002].

Thirty experiments are done on each feature set using each classifier. In each experiment, training sets are randomly chosen and the remaining sets are tested. The testing results consist of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In each category of the image complexity, the number of cover samples is approximately equal to the number of stego-samples, so the testing accuracy is calculated by (TP+TN) / (TP+TN+FP+FN). The average testing accuracy and the standard error of the thirty experiments are compared. Table 3-2 lists the testing results (mean values and standard deviations) on each feature set with the use of SVM, ADABOOST, NBC, and QDC. In each category of image complexity, the best testing accuracy is in bold. In the five categories of image complexity, all the highest testing results happen to the feature set of EHPCC. The results indicate that EHPCC is superior to its subset CF; CF is better than HCFHOM, A.HCFCOM, and C.A.HCFCOM; the detection performance of HOMMS is not good. The results in table 3-2 are consistent with the demonstration of the statistical significance in Fig. 3-3. Regarding the detection performance of these four learning classifiers, SVM and adaboost are better than NBC and QDC.

Since EPHCC is the best feature set, we compare the detection performance by applying DENFIS to EPHCC against the best testing values in table 3-2; the results are shown in table 3-3. On the average, DENFIS is better than SVM and adaboost.

### 3.5.4 Comparison of Feature Selections

Although EHPCC has the best detection performance, Fig. 3-3 shows that not all the features in EHPCC are good, not all the elements of HOMMS are useless. If we combine all the features listed in table 3-1, how to choose the features?

Since tables 3-2 and 3-3 show that DENFIS is better than SVM and adaboost and Sequential Forward Selection (SFS) is a classical approach in wrapper feature selections, we compare DENFIS based SFS (DENFIS-SFS) with SVMRFE and T-test. Fig. 3-4 plots the cross-validation detection performances under the feature dimension one to forty with the application of DENFIS and SVM to the feature selections: SVMRFE, DENFIS-SFS, and T-test. It shows that, while $\beta > 0.8$, by applying SVM to all the feature sets from feature dimension one to forty, it fails to detect the steganography; on the contrary, DENFIS works well. Fig. 3-4 indicates that, regarding the testing accuracy and the stability spanning over different image complexity, the classifier DENFIS outperforms SVM; the feature selection SVMRFE is superior to DENFIS-SFS and DENFIS-SFS is better than T-test; the combination of DENFIS with SVMRFE achieves the best detection performance.

Table 3-2 Testing results on the feature sets (mean value ± standard deviation, %). In each category of image complexity, the highest test accuracy is in bold. As β > 0.8, SVM fails to classify the HOMMS feature set.

| β | Classifier<br>Feature set<br>and testing accuracy | SVM | ADABOOST | NBC | QDC |
|---|---|---|---|---|---|
| < 0.4 | EHPCC | **91.8 ± 0.9** | 89.0 ± 1.0 | 76.0 ± 1.9 | 70.3 ± 1.8 |
| | CF | 85.9 ± 1.0 | 82.0 ± 1.2 | 77.0 ± 2.1 | 80.9 ± 1.7 |
| | HCFHOM | 60.9 ± 1.3 | 57.6 ± 1.5 | 57.5 ± 1.5 | 53.4 ± 1.0 |
| | HOMMS | 53.6 ± 1.0 | 50.6 ± 2.0 | 46.9 ± 1.7 | 42.1 ± 1.4 |
| | C.A.HCFCOM | 55.3 ± 0.6 | 54.3 ± 1.1 | 53.8 ± 1.1 | 55.4 ± 1.1 |
| | A.HCFCOM | 55.6 ± 0.9 | 55.4 ± 1.8 | 54.7 ± 1.4 | 55.5 ± 1.1 |
| 0.4-0.6 | EHPCC | **86.2 ± 0.6** | 79.9 ± 1.0 | 66.8 ± 0.9 | 65.1 ± 0.8 |
| | CF | 77.6 ± 0.4 | 72.2 ± 1.0 | 67.6 ± 1.3 | 70.6 ± 1.3 |
| | HCFHOM | 58.4 ± 0.6 | 56.6 ± 1.1 | 56.1 ± 0.9 | 54.5 ± 0.6 |
| | HOMMS | 48.8 ± 1.6 | 47.6 ± 1.0 | 47.1 ± 0.8 | 44.0 ± 1.5 |
| | C.A.HCFCOM | 58.1 ± 0.7 | 57.0 ± 1.5 | 57.8 ± 1.1 | 57.9 ± 0.8 |
| | A.HCFCOM | 57.3 ± 0.6 | 56.6 ± 0.9 | 56.8 ± 0.7 | 56.6 ± 0.6 |
| 0.6-0.8 | EHPCC | **73.7 ± 1.3** | 69.4 ± 1.2 | 61.4 ± 1.4 | 62.8 ± 0.9 |
| | CF | 66.7 ± 0.7 | 63.9 ± 1.2 | 62.1 ± 1.1 | 62.3 ± 1.2 |
| | HCFHOM | 57.6 ± 0.9 | 55.3 ± 1.1 | 54.2 ± 1.3 | 53.1 ± 0.7 |
| | HOMMS | 47.3 ± 0.7 | 43.7 ± 1.3 | 45.4 ± 1.2 | 40.6 ± 2.4 |
| | C.A.HCFCOM | 56.0 ± 1.1 | 56.4 ± 1.0 | 55.8 ± 1.0 | 56.2 ± 0.8 |
| | A.HCFCOM | 56.6 ± 0.6 | 54.9 ± 1.2 | 55.2 ± 1.1 | 55.5 ± 1.2 |
| 0.8-1 | EHPCC | **63.7 ± 1.0** | 63.0 ± 1.4 | 56.5 ± 1.2 | 61.4 ± 1.0 |
| | CF | 60.0 ± 1.0 | 57.4 ± 1.8 | 57.8 ± 1.5 | 57.5 ± 1.6 |
| | HCFHOM | 53.9 ± 1.2 | 52.0 ± 1.6 | 53.2 ± 1.4 | 51.7 ± 0.6 |
| | HOMMS | / | 42.0 ± 1.5 | 44.5 ± 0.8 | 41.6 ± 2.8 |
| | C.A.HCFCOM | 52.4 ± 0.7 | 52.6 ± 1.5 | 52.1 ± 1.3 | 53.1 ± 1.2 |
| | A.HCFCOM | 53.3 ± 1.0 | 50.3 ± 1.3 | 51.8 ± 1.2 | 50.8 ± 1.6 |
| >1 | EHPCC | 54.6 ± 0.2 | **61.3 ± 1.2** | 58.0 ± 1.2 | 60.0 ± 0.5 |
| | CF | 59.7 ± 1.7 | 58.9 ± 2.3 | 57.1 ± 1.5 | 58.4 ± 1.3 |
| | HCFHOM | 54.4 ± 0.8 | 52.7 ± 1.6 | 51.9 ± 1.7 | 53.2 ± 1.8 |
| | HOMMS | / | 46.7 ± 1.8 | 50.4 ± 1.4 | 43.1 ± 1.5 |
| | C.A.HCFCOM | 54.7 ± 0.5 | 52.7 ± 1.7 | 53.1 ± 1.4 | 54.4 ± 0.9 |
| | A.HCFCOM | 54.3 ± 0.3 | 51.2 ± 1.6 | 51.6 ± 2.0 | 53.5 ± 1.4 |

Table 3-3 Applying DENFIS to EHPCC vs. the best results in Table 2.

| β | DENFIS | BEST TESTING IN TABLE 2 |
|---|---|---|
| < 0.4 | **93.2 ± 1.1** | 91.8 ± 0.9 |
| 0.4 – 0.6 | **87.7 ± 1.2** | 86.2 ± 0.6 |
| 0.6 -0.8 | 72.6 ± 1.6 | **73.7 ± 1.3** |
| 0.8 - 1 | 62.5 ± 2.2 | **63.7 ± 1.0** |
| > 1 | **62.8 ± 1.8** | 61.3 ± 1.2 |

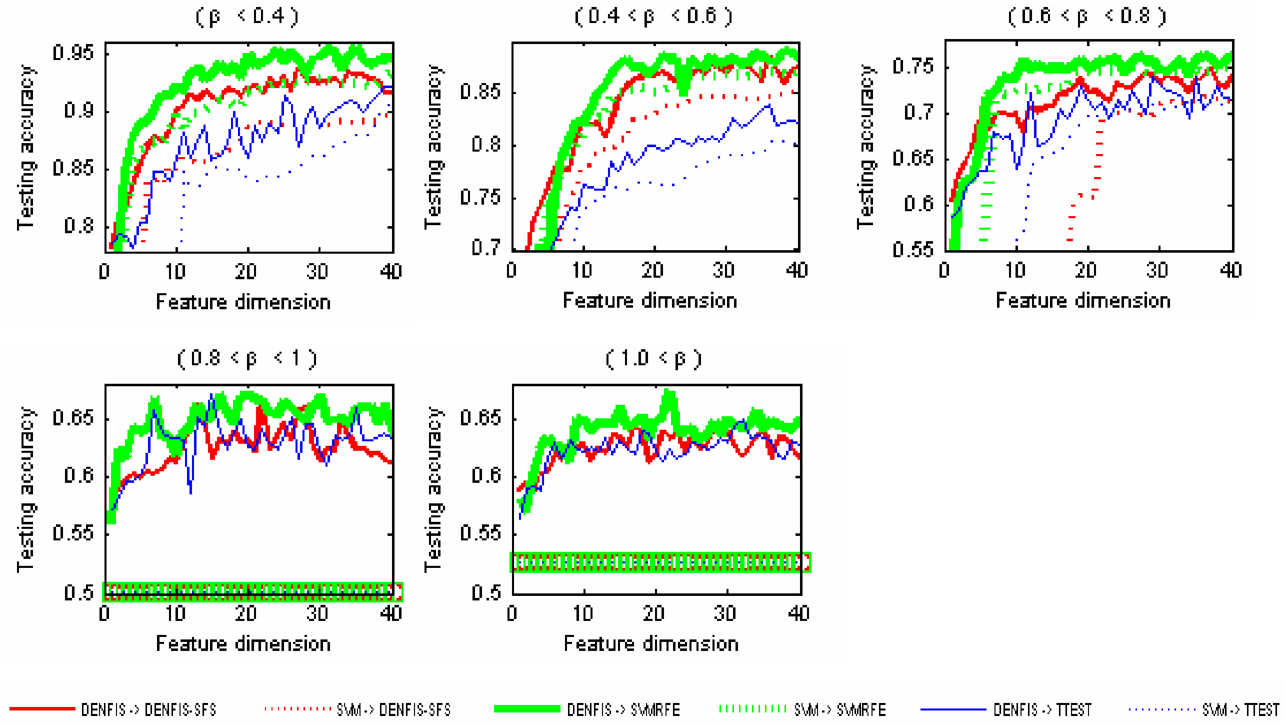


Fig. 3-4 The detection performance with the use of SVM and DENFIS to the feature selections: SVMRFE, DENFIS-SFS, and T-test. In the lower subfigures (0.8 < β < 1 and 1.0 < β), SVM fails to classify the testing sets of covers and stego-images.

### 3.6 Conclusions

In this chapter, a scheme of detecting LSB matching steganography in grayscale images is presented based on feature mining and pattern recognition techniques. Five types of features are extracted and several learning classifiers are applied. Experimental results indicate that the proposed feature set is better that other well-known feature sets including HCFHOM, HOMMS, A.HCFCOM, and C.A.HCFCOM. DENFIS is superior to adaboost, SVM, NBC, and QDC. To deal with the issue of feature selection in steganalysis, we compared three feature selections: SVMRFE, DENFIS-SFS, and T-test. SVMRFE performs the best. The learning classifier DENFIS combining with the feature selection of SVMRFE achieves the best detection performance.

The experimental results also show that image complexity is an important parameter for evaluating the steganalysis performance. At a certain information-hiding ratio, the detection performance is highly different in different image complexity. It is still very challenging in detecting the information-hiding behavior in the grayscale images with high complexity.

# CHAPTER 4: STEGANALYSIS OF TRANSFORM-HIDING STEGANOGRAPHY

## 4.1 Introduction

Transform-hiding steganography hides data in the coefficients of the transform domain such as DCT, DWT or DFT. JPEG image is one of the most popular media in Internet and it is easily used to carry hidden data and many information-hiding techniques/tools embed data in JPEG images; therefore, it's important for many purposes to design a reliable algorithm to decide whether a JPEG image found on the Internet carries hidden data or not.

There are a few methods for detecting JPEG steganography. HCFCOM and HOMMS are two well-known universal detectors which are described in the previous chapters, and they are suitable in detecting the information-hiding in JPEG images. Additionally, Fridrich *et al.* [Fridrich et al., 2003] presented a method to estimate the cover-image histogram from the stego-image. Another new feature-based steganalytic method for JPEG images was proposed and the features are calculated as an L1 norm of the difference between a specific macroscopic functional calculated from the stego-image and the same functional obtained from a decompressed, cropped, and recompressed stego-image [Fridrich, 2004]. Harmsen and Pearlman [Harmsen and Pearlman, 2004] implemented a detection scheme using only the indices of the quantized DCT coefficients in JPEG images.

In this chapter, a steganalysis scheme for JPEG images using polynomial fitting is presented. Many stegnographic systems in JPEG images modify the quantized DCT coefficients; as a result, the marginal density of the coefficients is affected. Based on this observation and concern, polynomial fitting is designed to fit the logarithmic transform domain of the marginal density, and the errors between the histogram and the fitting curve are extracted as the detector. Classification techniques are utilized to recognize the different types of the steganograms and the covers. In this chapter, an evolutionary neuro-fuzzy inference system is introduced to estimate the information-hiding length in the steganograms based on the detector. Experimental results indicate that this method is very successful in detecting the information-hiding types and the information-hiding length in the imbalance multi-class environment which consists of plenty of covers, and the JPEG steganograms produced by CryptoBola, F5, and JPHS information-hiding systems.

In the following part, JPEG compression and the information-hiding is introduced, and the detector of the errors of the Generalized Gaussian Distribution (GGD) model of the quantized DCT coefficients and the polynomial fitting is designed [Liu, Sung, Xu and Venkataramana, 2006], then the experiments and the results are demonstrated.

## 4.2 JPEG Compression and Information-hiding

JPEG is the image compression standard developed by the Joint Photographic Experts Group (official name ITU-T T.81, ISO/IEC IS 10918-1). In practice, JPEG is most often used to compress 24-bit color or 8-bit grayscale images. In 24-bit color images, each numeric value that describes the color of a pixel in a 24-bit color image actually breaks down into three values that define the exact color. There are two ways to define this set of three color values. Most of the computer-literate are familiar with the "RGB" color description scheme, where each pixel value is a set of by three numbers giving the red, green, and blue color value. For example, in RGB, each 24-bit value breaks down into three 8-bit values, each giving the intensity of red, green, and blue in a scale from 0 to 255. In the "luminance-chrominance" or YCbCr scheme, used in traditional US analog color TV, a pixel value is given by its grayscale brightness level, or "luminance", and by a color value, or "chrominance". Chrominance actually amounts to two values, one that describes the "hue", or specific color within a linear range of colors, and the other that describes the "saturation", or intensity of the color. The luminance information contains most of the detail perceived by the human eye, while the overlying chrominance color information can be fuzzy without causing any serious image degradation. JPEG compression applies luminance-chrominance scheme because it offers greater possibilities for compression. For example, compression can be increased by only sampling every other horizontal and vertical pixel in a chrominance block, which cuts the number of chrominance bits to a fourth. This is known as "horizontal and vertical decimation" using a factor of 2, and results in one decimated 8x8 chrominance block for every four luminance blocks. JPEG divides up each of the three YCbCr color planes into 8 by 8 pixel blocks, and then calculates the discrete cosine transform (DCT) of each block. A quantizer rounds off the DCT coefficients according to the quantization matrix. This step produces the "lossy" nature of JPEG, but allows for large compression ratios.

JPEG compression technique uses a variable length code on these coefficients, and then writes the compressed data stream to an output file.

Generally, many steganographic systems in JPEG images implement information-hiding by modifying the quantized DCT coefficients, e.g., JPEG-JSteg sequentially replaces the least-significant bit of DCT coefficients with the message's data, but it is easy to detect [Zhang and Ping, 2003]. Instead of replacing the least-significant bit of DCT coefficient with message data, F5 decrements its absolute value in a process called matrix encoding [Westfeld, 2001].

## 4.3 Detector of Errors of Polynomial Fitting (EPF)

As mentioned in chapter 2, several papers describe the Generalized Gaussian Distribution (GGD) model in transform domains, such as DCT, DFT, or DWT [Sharifi and Leon-Garcia, 1995]. The marginal density of DCT coefficients may be achieved by adaptively varying two parameters of the GGD, which is defined as follows:

$$\rho(x;\alpha,\beta)=\frac{\beta}{2\alpha\Gamma(1/\beta)}\exp\{-(|x|/\alpha)^{\beta}\} \tag{4-1}$$

Where $\Gamma(\bullet)$ is the gamma function, $\alpha$ models the width of the probability distribution function (PDF) peak and $\beta$ is inversely proportional to the decreasing rate of the peak. $\alpha$ is referred to as the scale parameter while $\beta$ is called the shape parameter. The GGD model contains the Gaussian and Laplacian PDFs as special cases, using $\beta = 2$ and $\beta = 1$, respectively.

For the quantized JPEG DCT coefficients, the values of *x* in (1) are the discrete values, 0, 1, -1, 2, -2, 3, -3, etc. The marginal density of the quantized JPEG DCT coefficients, *h*(*x*), can be approximately modeled as follows:

$$h(x) \ = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\{-(|x|/\alpha)^{\beta}\},\ x = 0,\ 1,\ \text{-}1,\ 2,\ \text{-}2,\ \ldots \tag{4-2}$$

Applying logarithmic to (2), in the case of $x > 0$,

$$f(x) = \log\{h(x)\} = \log\{\frac{\beta}{2\alpha\Gamma(1/\beta)}\} - (|x|/\alpha)^{\beta} = A - B \cdot x^{\beta} \tag{4-3}$$

Where $A = \log\{\frac{\beta}{2\alpha\Gamma(1/\beta)}\};\ B = \alpha^{-\beta}$ .

A Taylor series to (3),

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots \tag{4-4}$$

When $a$ is set to 0, *f(x)* can be approximately represented by the $n^{\text{th}}$ polynomial series. Considering the computational complexity of the Taylor series, we denote $p_n(x)$ as the $n^{\text{th}}$ polynomial that fits function *f(x)* best in a least-square sense.

$$p_n(x) = p(1) \bullet x^n + p(2) \bullet x^{n-1} + \ldots + p(n) \bullet x + p(n+1) \tag{4-5}$$

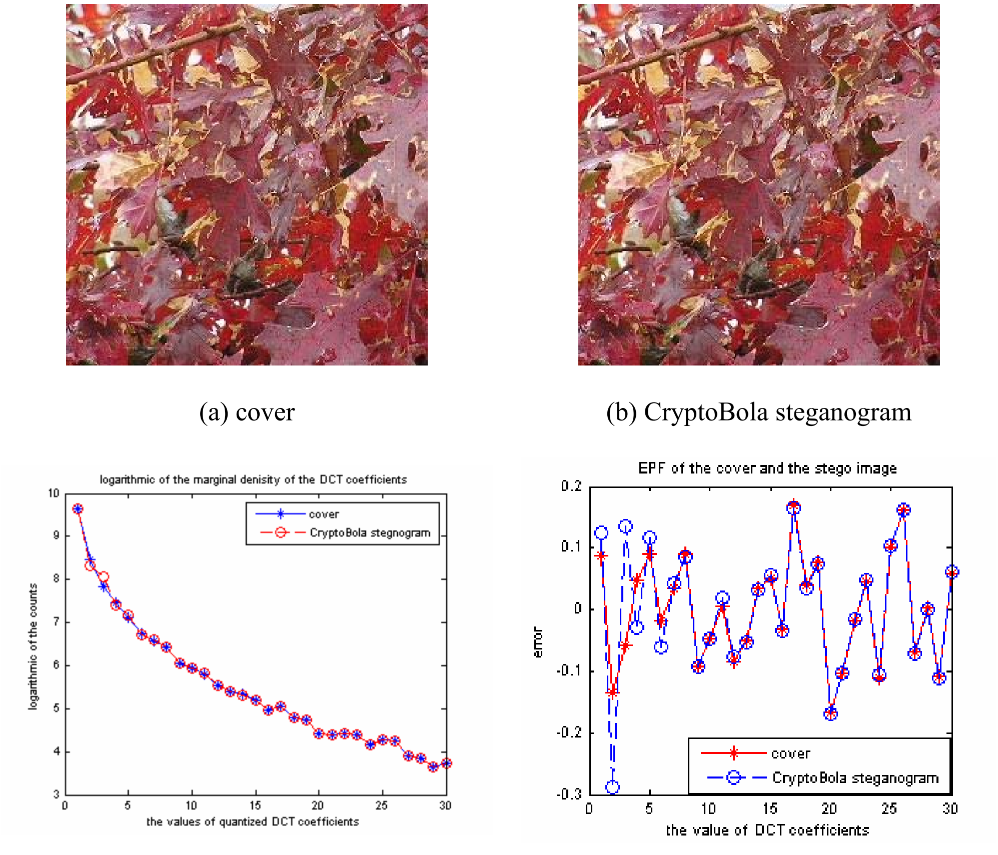
The error at any value for *x* is defined as:

$$R_n(x) = f(x) - p_n(x) \tag{4-6}$$

We call the measure $R_n(x)$ Errors of Polynomial Fitting (EPF).

Generally, most JPEG steganographic systems modify the quantized DCT coefficients. It may affect the marginal density of the DCT coefficients and the distribution may deviate from the GGD. As a result, some EPF errors of the steganograms will differ from those of the untouched JPEG images. Hence, the presence of hidden data in these JPEG steganography may be caught according to the statistics of the $R_n(x)$ .

Fig. 4-1(a) lists a JPEG cover of 18232 bytes and Fig. 4-1(b) shows the CryptoBola JPEG steganogram of 18202 bytes wherein a text file of 682 bytes is hidden. The hidden text file is not shown here. Fig. 4-1(c) shows the logarithmic of the marginal densities of the quantized DCT coefficients and Fig 4-1(d) demonstrates the EPF. Fig. 4-1(c) indicates that the marginal densities of the DCT coefficients are different between the cover and the steganogram, which results in the difference of the EPF (Fig. 4-1(d)).



(a) cover

(b) CryptoBola steganogram

(c)

(d)

Fig. 4-1 A JPEG cover (a), the steganogram (b), the logarithmic of the histogram of the DCT coefficients (c), and the EPFs (d).

## 4.4 Experimental Results

The original images are TIFF raw format digital pictures taken during 2003 to 2005. These images are 24-bit, 640×480 pixels, lossless true color and never compressed. We cropped the original images into 256×256 pixels in order to get rid of the low complexity parts. After that, we converted the cropped images into JPEG and the quality is 75 (the default quality). These JPEG images as well as other JPEG images collected during 2002 to 2003 are the original covers. The following three different information hiding techniques are adopted:

1. CryptoBola JPEG. It determines which parts (bits) of the JPEG-encoded data play the least significant role in the reproduction of the image, and replace those bits with the bits of the cipher text. CryptoBola is available at http://www.cryptobola.com/.
2. F5 algorithm [Westfeld, 2001]. This algorithm F5 withstands visual and statistical attacks, yet it still offers a large steganographic capacity. F5 implements matrix encoding to improve the efficiency of embedding. Thus it reduces the number of necessary changes. F5 employs permutative straddling to uniformly spread out the changes over the whole steganogram.
3. JPHS (JPHIDE and JPSEEK). The design objective was not simply to hide a file but rather to do this in such a way that it is impossible to prove that the host file contains a hidden file. Given a typical visual image, a low insertion rate (under 5%) and the absence of the original file, it is not possible to conclude with any worthwhile certainty that the host file contains inserted data. As the insertion

percentage increases the statistical nature of the jpeg coefficients differs from "normal" to the extent that it raises suspicion. JPHS for Windows (JPWIN) is available at: http://digitalforensics.champlain.edu/download/jphs_05.zip/.

In our experiments, we apply the sixth polynomial that fits the logarithmic of the histogram of the absolute values of the quantized DCT coefficients in the luminance component, and the error between the logarithmic of the histogram and the polynomial fit $R_6(n)$ ($n$ = 1, 2, …30) are extracted. Additionally, the measures of HCFCOM and HOMMS are extracted for comparison. Adaboost and SVM are applied to different feature sets. We perform each experiment 30 times. In each time, the training samples are randomly chosen and the remaining samples are tested for validation. The ratio of training to test samples is 1:1. The mean values and standard deviation of the test accuracy in the 30 times are compared.

Table 4-1 Detection performance (mean testing accuracy ± standard deviation, %) on different feature sets in binary class environment (cover and the steganogram)

| Hiding method<br>Classifier<br>Feature set | CryptoBola | | F5 | | JPHS | |
|---|---|---|---|---|---|---|
| | Adaboost | SVM | Adaboost | SVM | Adaboost | SVM |
| $\{R_6(n)\|n = 1,2, \ldots, 5\}$ | 100 ± 0 | 100 ± 0 | 95.6 ± 0.7 | 96.5 ± 0.7 | 85.8 ± 2.3 | 83.6 ± 2.3 |
| $\{R_6(n)\|n = 1,2, \ldots, 10\}$ | 100 ± 0 | 99.9 ± 0.1 | 95.5 ± 0.8 | 95.6 ± 0.8 | 87.7 ± 1.7 | 86.2 ± 2.8 |
| $\{R_6(n)\|n = 1,2, \ldots, 20\}$ | 99.9 ± 0.1 | 99.8 ± 0.1 | 94.1 ± 0.7 | 95.0 ± 0.8 | 87.5 ± 2.0 | 83.4 ± 2.2 |
| HCFCOM | 56.4 ± 1.6 | 53.1 ± 2.2 | 59.7 ± 1.8 | 55.8 ± 2.1 | 62.7 ± 2.7 | 66.9 ± 3.6 |
| HOMMS | 73.6 ± 1.7 | 50.0 ± 0.1 | 77.2 ± 1.7 | 50 ± 0 | 81.2 ± 2.7 | 50 ± 0 |

Table 4-1 lists the mean testing accuracy and the standard deviation using adaboost and SVM with different feature sets, showing that the detection performance of EPF is superior to those of HCFCOM and HOMMS. Fig. 4-2 plots the ROC curves of the detection performance on different feature sets with the use of adaboost. Table 4-1 and Fig. 4-2 indicate that EPF is the best detector in the steganalysis of the three types of JPEG steganography. Fig. 4-2(a) indicates that in steganalaysis of CryptoBola steganography, the area below the EPF curve (EPF curve is overlapped with the *x*-axis) is zero, which means that the detection performance on EPF is perfect, there is no error in classification of covers and steganograms. However, the detection performances on HCFCOM and HOMMS are not good. Fig. 4-2(b) and Fig. 4-2(c) also indicate that the detection performances on EPF are the best, and those on HCFCOM and HOMMS are not so good.

Table 4-2 lists the testing results to the detector of EPF with the use of One-Against-All decomposition for Support Vector Machine (OAASVM) [Schlesinger and Hlavac, 2002; Vapnik 1998]. Table 4-2 indicates that, in the multi-class JPEG images, by applying OAASVM to EPF, the correction prediction for covers, cryptobola and F5 steganograms is very successful; but the discrimination between covers and JPHS steganograms is not so good. To obtain a better prediction for covers and JPHS steganograms, we apply adaboost to the EPF features. Table 3 gives the testing results. Obliviously, just regarding the classification of covers and the JPHS on the EPF features, adaboost is superior to OAASVM.

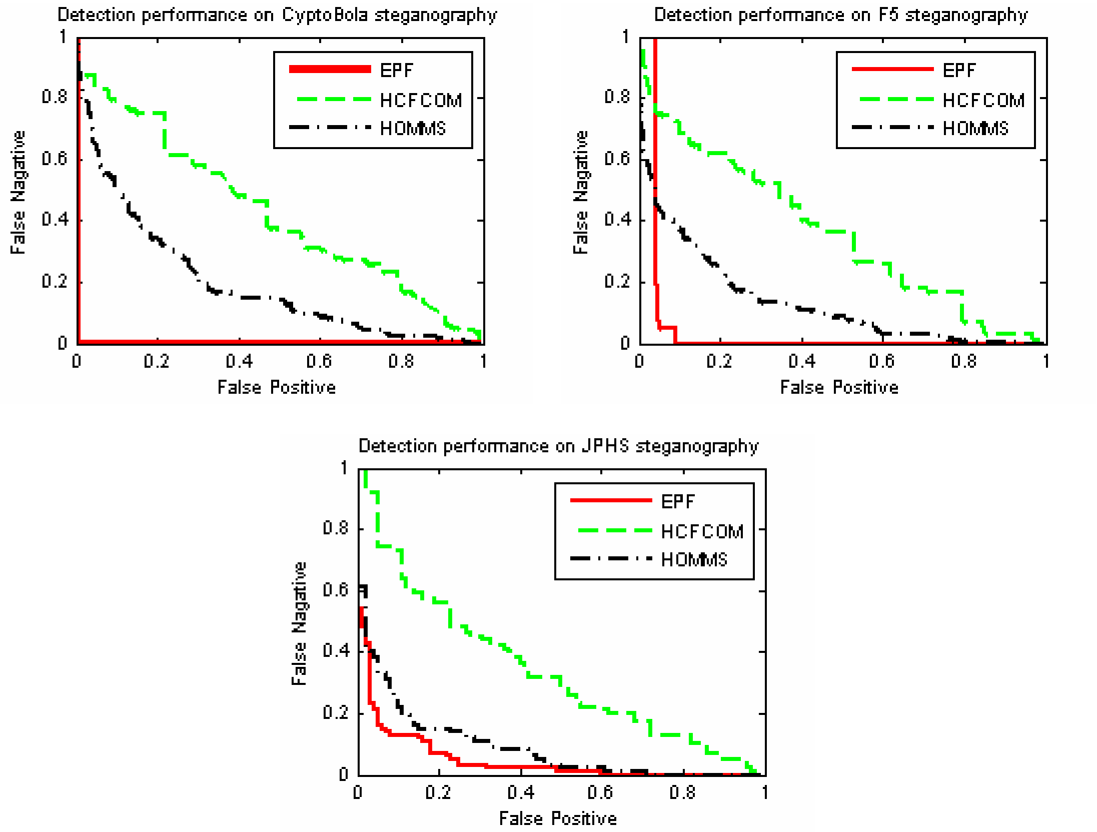


Fig. 4-2 ROC curves in the steganalysis of the three types of steganography

Table 4-2 The multi-class prediction in the multi-class JPEG images with the use of OAASVM

| Multi-class prediction<br>Multi-class testing sets | | Cover | CryptoBola | F5 | JPHS | Correction prediction in the multi-class |
|---|---|---|---|---|---|---|
| Cover | 10000 | 9967 | 3 | 30 | 0 | 99.7% |
| CryptoBola | 800 | 3 | 796 | 1 | 0 | 99.5% |
| F5 | 800 | 75 | 1 | 724 | 0 | 90.5% |
| JPHS | 400 | 389 | 8 | 0 | 3 | 0.8% |

Table 4-3 The prediction between covers and JPHS steganograms with the use of Adaboost

| Prediction<br>Testing sets | | Cover | JPHS | Correction prediction |
|---|---|---|---|---|
| Cover | 10000 | 9926 | 74 | 99.3% |
| JPHS | 400 | 175 | 225 | 56.3% |

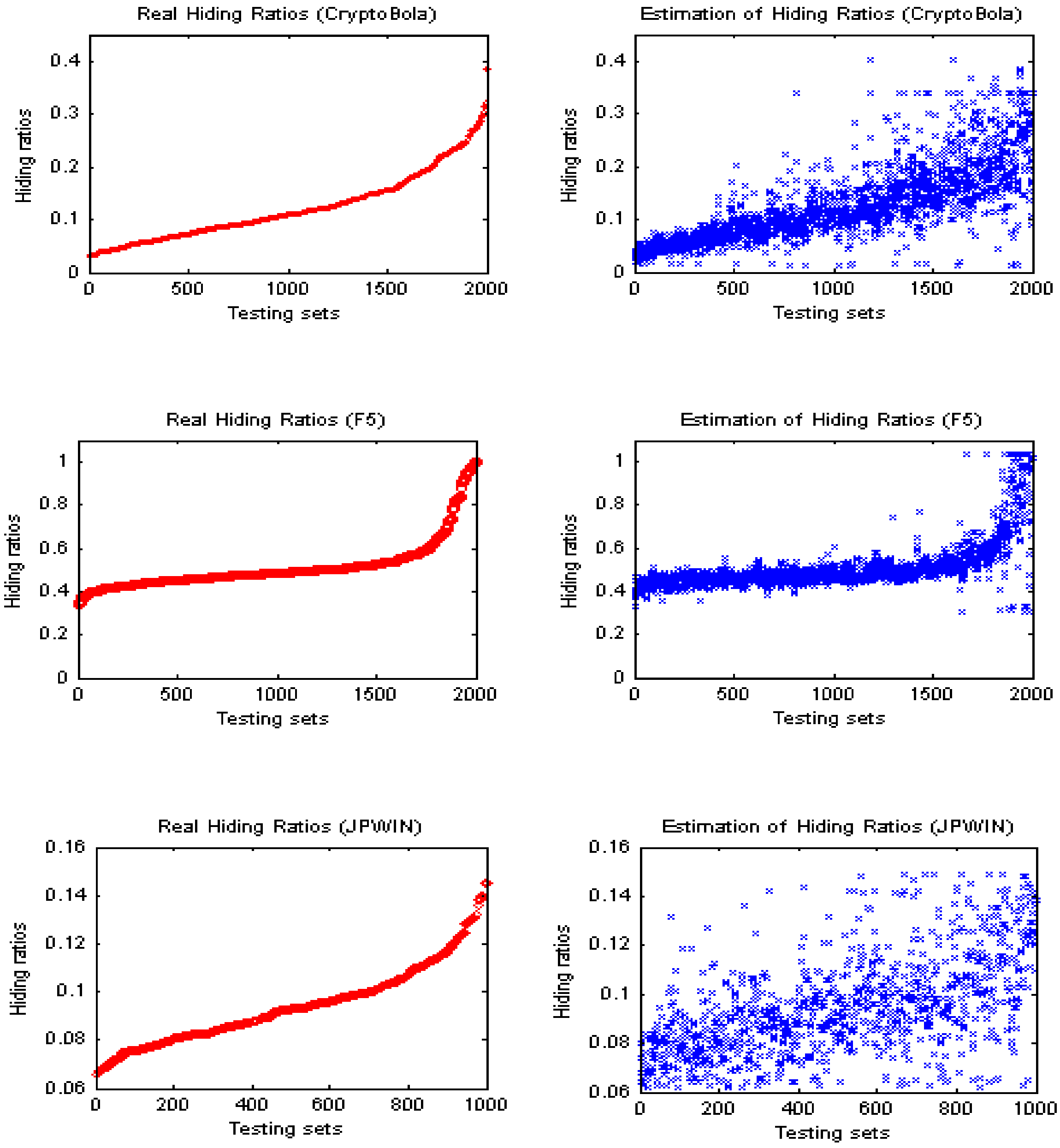


Fig. 4-3 The real distribution of the information-hiding ratios (left) and the prediction (right) with the use of DENFIS.

We also apply DENFIS to predict the information-hiding length (ratio) in the JPEG steganograms. Here we specially measure the information-hiding ratio as the ratio of the modification-length of the non-zero quantized DCT coefficients to the length of the non-zero quantized DCT coefficients.

Fig. 4-3 shows the real distribution of the information-hiding ratio (left) and the prediction (right) on the three JPEG stego-images. The prediction of the hiding ratio is denoted as *pr*, the real hiding ratio is denoted as *rr.* We adopt the following measure EP to evaluate the error of the prediction.

$$\mathrm{EP} = \mathrm{abs}(pr - rr) / rr * 100\% \tag{4-7}$$

Table 4-4 gives the mean values of EP and the standard errors in the three JPEG steganograms.

Table 4-4 Mean values and standard errors of the EPFs.

| | CryptoBola | F5 | JPHS (JPWIN) |
|---|---|---|---|
| mean(EP) / std(EP), % | 6.82 / 7.7 | 22.4 / 22.9 | 13.1 / 11.3 |

## 4.5 Conclusions and Future Work

In this chapter we propose a scheme of steganalysis of JPEG images. We extract the errors between the logarithmic of the marginal density of the quantized DCT coefficients and the polynomial fitting as the detector, and apply several computational techniques to the detection. Results show that, designed method is successful in detecting the presence of hidden data in the JPEG steganograms produced by CryptoBola, F5, and JPHS. It is superior to the well-known methods of HCFCOM and HOMMS. We apply OAASVM, adaboost, and DENFIS to the EPFs of the imbalance multi-class JPEG images. Results indicate that our method is successful in detecting the information-hiding types and the information-hiding length.

Future work includes improving and expanding method to detect the location of information-hiding by combining the features proposed by Fridrich [Fridrich, 2004] and extracting the payload without the prior knowledge of the information-hiding techniques.

# CHAPTER 5: INTRODUCTION TO BIOINFORMATICS

## 5.1 Bioinformatics in Brief

Bioinformatics derives knowledge from computer analysis of biological data and is the intersection of multiple science fields including molecular biology, computer science, statistics, etc. There are various definitions of bioinformatics on the Web.

Bioinformatics definition by bioinformatics definition Committee, National Institute of Mental Health released on July 17, 2000 (source: http://www.bisti.nih.gov/ )

"The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

***Bioinformatics:*** Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

***Computational Biology:*** The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."

The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as "***Bioinformatics*** is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."

Major research efforts in bioinformatics include sequence analysis, genome annotation, computational evolutionary biology, measuring biodiversity, analysis of gene expression, analysis of regulation, analysis of protein expression, analysis of mutation in cancer, prediction of protein structure, comparative genomics, and high-throughput image analysis, etc. [ http://en.wikipedia.org/wiki/Bioinformatics]

## 5.2 Introduction to Microarays and SNPs

### Single Nucleotide Polymorphisms (SNPs)

A **Single Nucleotide Polymorphism** or **SNP** (pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGC**C**TA to AAGC**T**TA, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T. SNPs typically have three genotypes, denoted generically AA, Aa and aa. In the example above, the three genotypes would be CC, CT and TT. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. These changes may cause disease, and may affect how a person reacts to bacteria, viruses, drugs, and other substances. For example, Sickle cell anemia (SCA) is the most common inherited blood disorder in the United States, affecting about 72,000 Americans or 1 in 500 African Americans. SCA is an autosomal recessive disease caused by a point mutation (SNP) in the *hemoglobin beta gene (HBB)* found in region 15.5 on the short arm (p) of chromosome 11 [*genes and diseases*, http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=gnd].

Additionally, the following phenomena are common in our life.

1) One man who drinks alcohol and smokes cigarettes lives to age 90 without getting liver or lung cancer; another man who smokes and drinks the same amount gets cancer at age 60; the third one who does not smoke and drink gets cancer at age 55.

2) One woman's breast cancer responds to chemotherapy, and her tumor shrinks; another woman's breast cancer shows no change after the same treatment.

How do we explain these differences? SNPs in the human genome may be the solutions. The human genome is the complete set of instructions for life. Except for red blood cells, which have no nucleus, the human genome is located in the nucleus of every cell in the body. There are 22 pairs of chromosomes and one pair of sex chromosomes. Chromosomes are made of deoxyribonucleic acid (DNA), which contains only four chemical bases or building blocks: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). There are roughly 3.2 billion chemical bases (A, T, C, G) in the human genome. Each DNA molecule is made up of two long complementary (related) strands, or "double helix" and A always pairs with T, and C with G, the order on one strand dictates the order on the other. Only about 3 percent of the human genome is actually used as the set of instructions and these regions are called coding regions and scattered throughout the chromosomes.

A coding region contains genes. A gene is a unique DNA sequence within a chromosome that ultimately directs the building of a specific protein with a specific function. Close to each gene is a "regulatory" sequence of DNA, which is able to turn the gene "on" or "off." There are at least 35,000 genes in the human genome, and there may be more. There is no function for most of the remaining 97 percent of the genome. These regions are called noncoding regions. An amazing aspect of the human genome is that there is so little variation in the DNA sequence when the genome of one person is compared to that of another. Of the 3.2 billion bases, roughly 99.9 percent are the same between any two people. It is the variation in the remaining tiny fraction of the genome, 0.1 percent--roughly several million bases--that makes a person unique. This small amount of variation determines attributes such as how a person looks, or the diseases he or she develops. Most variations in the human genome have no known effect at all because they occur in noncoding regions of the DNA. In addition, there are some changes that do occur in coding and regulatory regions, yet they have no known effect. All these are silent variations. Some of the variations that occur in the coding and regulatory regions of genes have "harmless" effects. They can, for example, change the way a person "looks." Some people have blue eyes, others brown; some are tall, others short; and some faces are oval, others round.  Other variations in coding regions are harmless because they occur in regions of a gene that do not affect the function of the protein made.

There are a group of variations in coding and regulatory regions that result in harmful effects. These are called mutations. They cause disease because changes in the genome's instructions alter the functions of important proteins that are needed for health. For example, diabetes, cancer, heart disease, Huntington's disease, and hemophilia all result from variations that cause harmful effects. In a "simple" disease such as hemophilia, variation in one gene is sufficient to cause disease symptoms. By contrast, in a "complex" disease like cancer, symptoms are seen only after many variations have occurred in different genes in the same cell. Finally, there are genetic variations that have "latent" effects. These variations, found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under certain conditions. Such changes may eventually cause some people to be at higher risk for cancer, but only after exposure to certain environmental agents. They may also explain why one person responds to a drug treatment while another does not. Here is part of the genome from two people who are both smokers and drinkers, but only one of them gets cancer. The zoom into the chromosomes of these two men shows just a sampling of the differences in variation that are responsible for their individual cancer risk. The variations themselves do not cause cancer. They only affect each person's susceptibility to tobacco smoke and alcohol after exposure.

SNPs are scattered throughout the genome and are found in both coding AND noncoding regions. SNPs can cause silent, harmless, harmful, or latent effects. They occur with a very high frequency, with estimates ranging from about 1 in 1000 bases to 1 in 100 to 300 bases. This means that there could be millions of SNPs in each human genome. The abundance of SNPs and the ease with which they can be measured make these genetic variations significant. Most SNPs occur in non-coding regions and do not alter genes. Scientists are finding that some of these SNPs have a useful function. If a SNP is frequently found close to a particular gene, it acts as a marker for that gene. The remaining SNPs occur in coding regions. They could alter the **protein** made by that coding region, which in turn could influence a person's health [http://www.nci.nih.gov/cancertopics/understandingcancer/geneticvariation].

### Microarrays

In the past several years, a new technology, called DNA **microarray**, has attracted tremendous interests among biologists. This technology promises to monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of genes simultaneously.

***Microarray*** is *a* 2D array, typically on a glass, filter, or silicon wafer, upon which genes or gene fragments are deposited or synthesized in a predetermined spatial order allowing them to be made available as probes in a high-throughput, parallel manner. Microarrays include different kinds of biological assays: DNA microarrays, protein microarrays, tissue microarrays, transfection microarrays, chemical compound microarrys, and antibody microarrays. A **DNA microarray** (also commonly known as *gene chip*, *DNA chip*, *genome chip* or *gene array*) is a collection of microscopic DNA spots, arrayed on a solid surface by covalent attachment to chemically suitable matrices. An array is an orderly arrangement of samples. It provides a medium for matching known and unknown DNA samples based on base-pairing rules (i.e., A-T and G-C for DNA; A-U and G-C for RNA) and automating the process of identifying the unknowns [Simon *et al.*, 2003; Muller and Nicolau, 2005].

DNA **microarray**, or DNA chips are fabricated by high-speed robotics, generally on glass but sometimes on nylon substrates, for which probes with known identity are used to determine complementary binding, thus allowing massively parallel gene expression and gene discovery studies. An experiment with a single DNA chip can provide researchers information on thousands of genes simultaneously [Lander *et al*., 1999, Allison, 2005].

There are three types of microarrays, two are genomic and the other is “transcriptomic”, which measures mRNA levels. The first one is called microarry expression analysis, which determines the gene expression level, or volume. And the arrays in this type of analysis, so-called “expression chips”, can are used in drug development, drug response, and therapy development. The second called microarray Comparative Genomic Hybridization (CGH) is applied to look for genomic gains and losses or for a change in the number of copies of a particular gene involved in a disease state. The third one is used to detect mutations or polymorphisms in a gene sequences, the target, or immobilized DNA including single nucleotide polymorphism (SNP). [http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html].

The microarray (DNA chip) technology is having a significant impact on genomics study. Many fields, including drug discovery and toxicological research, etc., will certainly benefit from the use of DNA microarray technology. For example, if a certain gene is over-expressed in a particular cancer, expression chips can be used to see if a new drug will reduce over-expression and force the cancer into remission. In response to infection, certain cell types will express sets of genes and synthesize certain proteins that respond to the stress. Messenger RNA (mRNA) is like a photocopy of a blueprint that is used in the shop to build a specific type of protein. In a microarray, we can attach sequences from a range of genes to a glass slide in a series of dots, and then bind the mRNA extracted from a population of cells and measure how much binds to each dot. That gives us a snapshot of which genes are being expressed at any given time. Compare the patterns for mRNA from, for example, normal breast tissue and from a breast tumor, and you can identify proteins that are only present in the tumor. Those proteins are potential targets for cancer treatments, vaccines, and other therapeutics. Other applications of microarrays include tumor classification, risk assessement, and prognosis prediction, drug development, therapy development, and tracking disease progression, etc. The more details can be found in the source [http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html].

Since microarray can be used to examine the expression of thousands of genes simultaneously, absolutely, it promises to revolutionize the way scientists examine gene expression and represent an important and necessary first step in our understanding and cataloging of the human genome. Microarray data may contain high variables of the genes, it is very important and challenging to mine the critical or related genes from the microarray data and construct the association between the data and the phenotypes. Under my advisor's direction, I focused my dissertation research in bioinformatics on microarray analysis and tagging SNP selection. The reminder chapters are organized as follows. Chapter 6 presents the algorithms of recursive feature addition and lagging prediction peephole optimization to improve the classifications of microarray data analysis. Chapter 7 expands the algorithm of recursive feature addition to tagging SNP selection and introduces a method of SNP selection by calculating the support vector weights and the idea of recursive feature addition.

# CHAPTER 6: MICROARRAY GENE EXPRESSION ANALYSIS

## 6.1 Related Work in Microarray Analysis

Microarrays are capable of profiling the gene expression patterns of tens of thousands of genes in a single experiment. One of the key challenges of microarray studies is to derive biological insights from the unprecedented quantities of data on gene expression patterns. Partitioning genes into closely related groups across time with clustering techniques and classification of the patients based on the selected gene signatures have become two main tracks of practically all analyses of microarray data in the past decade [Quackenbush, 2001; Hand and Heard, 2005; Segal *et al*., 2005; Tjaden, 2006; Qin, 2006; Sha *et al*., 2006]. Statistical modeling and inference problems with sample sizes substantially smaller than the number of available features/genes are challenging, which is known as the "large *p* small *n* problem". Moreover, exploiting information redundancy from highly correlated genes/features may potentially reduce the efforts in terms of time and cost for genetic studies in human genetic research. The two fundamental questions and challenges of the high dimensional gene data are how many genes is enough to provide good prediction performance of disease status and how to determine the optimal final gene set that are best for predictions and classifications.

To address the "curse of dimensionality" problem, generally, such efforts can be grouped into three categories: filtering, wrapper, and embedded methods. Filtering methods select feature subsets independently from the learning classifiers and do not incorporate learning [Newton et al., 2001; Long et al., 2001; Bo and Jonassen, 2002; Yu and Chen, 2005]. A weakness of filtering methods is that they only consider the individual features in isolation and ignore the possible interaction among them. Yet, the combination of these features may have a combined effect that does not necessarily follow from the individual performances of features in the group [Pavlidis and Noble, 2001]. One of the consequences of filtering methods is that we may end up with many highly correlated features/genes with highly redundant information that worsens the classification and prediction performance. If there is a limit on the number of features to be chosen, we may not be able to include all informative features.

To address this problem in filtering methods, wrapper methods wrap around a particular learning algorithm that can assess the selected feature subsets in terms of the estimated classification errors and then build the final classifier [Inza *et al*., 2002]. Wrapper methods use a learning machine to measure the quality of subsets of features. One of the recent well-known wrapper methods for feature/gene selection is Support Vector Machine Recursive Feature Elimination, which refines the optimum feature set by using Support Vector Machine [Guyon *et al*., 2002]. The idea of SVMRFE is that the orientation of the separating hyper-plane found by the SVM can be used to select informative features: if the plane is orthogonal to a particular feature dimension, then that feature is informative, and vice versa.

Wrapper methods can notably reduce the number of features and significantly improve the classification accuracy [Monari and Dreyfus, 2000; Rivals and Personnaz, 2003].

However, wrapper methods have the drawback of high computational cost. With much better computational efficiency and similar performance to wrapper methods, a relatively new class of approaches for feature selection called “embedded methods” has become available in the literature. Embedded methods process feature selection simultaneously with the learning classifier, therefore they can incorporate knowledge about the structure of the classification. LASSO proposed by Tibshirani [Tishirani, 1996, 1997]; logic regression with the regularized Laplacian prior [Krishanpuram *et al.*, 2005]; and Bayesian regularized neural network with automatic relevance determination [Liang and Kelemen, 2005] are examples of embedded techniques.

Combining the sequential forward selection (SFS) and sequential floating forward selection (SFFS) with LS (Least Squares) Bound measure, Zhou and Mao proposed SFS-LS bound and SFFS-LS bound algorithms for optimal gene selection [Zhou and Mao, 2005]. To improve the classification of microarray gene expression data, another two gene selection methods were proposed, one is leave-one-out calculation sequential forward selection (LOOCSFS) algorithm, and the other is the gradient based leave-one-out gene selection (GLGS) algorithm [Tang *et al.*, 2006]. Recently, Diaz-Uriarte and de Andres presented a new method for gene selection that uses random forest [Diaz-Uriarte and de Andres, 200]. The main advantage of this method is that it returns very small sets of genes that retain a high predictive accuracy. The algorithms are publicized in the R package of varSelRF.

In this chapter, a scheme of Recursive Feature Addition (RFA) is presented to deal with redundancy issues and to improve the classification accuracy [Liu and Sung, 2006]. The recursive procedure is based on the supervised learning with selected classifier and the statistical similarity measures between the chosen genes and the candidates. We compare RFA with above SVMRFE, LOOCSFS, GLGS, SFS-LSbound, SFFS-LSbound, and T-test using six benchmark microarray gene expression datasets. Moreover, we propose a new algorithm, called Lagging Prediction Peephole Optimization to choose the final optimal feature/gene set for improve the classification. We compared our LPPO to random strategy under the best training classification and also LPPO with RFA to the popular gene selection method with the use of RF using six benchmark datasets.

## 6.2 Recursive Feature Addition for Gene Selection

### 6.2.1 Supervised Recursive Learning

The method of recursive feature addition is based on supervised learning and statistical similarity measures between the chosen genes and the candidates. This new approach is an embedded method and is presented as follows:

1. Each individual gene is selected with supervised learning, and the gene with the highest classification accuracy is chosen as the most important feature, and the first element of the feature set. If multiple genes achieve the same highest classification accuracy, the lowest *p*-value measured by test-statistics (e.g., score test), is the target of

the first element. At this point the chosen feature set, $G_1$, consists of the first feature, $g_1$, which corresponds to feature dimension one.

2. The $N+1$ dimension feature set, $G_{N+1} = \{g_1, g_2, \ldots, g_N, g_{N+1}\}$ is produced by adding $g_{N+1}$ to the $N$ dimension feature set, $G_N = \{g_1, g_2, \ldots, g_N\}$. The choice of $g_{N+1}$ is described as follows:

Add each gene $g_i$ ($i \neq 1 \quad 2, \ldots, N$) outside of $G_N$ to $G_N$ and record the classification accuracy of the feature set $G_N + \{g_i\}$. The $g_c$ ($g_c \notin G_N$) corresponding to the highest classification accuracy is marked and put into the set of candidates, $C$. Generally, the set of candidates consists of multiple genes because of the high dimension of microarray data, but only one gene in $C$ will be chosen.

### 6.2.2 Candidate Feature Addition

To obtain a more informative and least redundant set, two strategies are designed for choosing $g_{N+1}$ by measuring the statistical similarity between the chosen genes and candidates. Here we apply Pearson's correlation coefficient [Tan *et al.*, 2005] between the chosen gene $g_n$ ($g_n \in G_N$, $n = 1, 2, \ldots, N$) and the candidate $g_c$ ($g_c \in C$, $c = 1, 2 \ldots m$; $m$ is the number of the elements in $C$) to measure the similarity.

In the first strategy, the Sum of the square of the Correlation (SC) is calculated to measure the similarity and is defined as follows:

$$\mathrm{SC}(g_c) = \sum_{n=1}^{N} \mathrm{cor}^2(g_c, g_n), \ \ n = 1, 2 \ldots N \qquad (6\text{-}1)$$

where, $g_c \in C, g_n \in G_N$.

The selection of $g_{N+1}$ follows the qualification that the SC value is the minimum:

$$\{g_{N+1} \mid g_{N+1} \in \mathrm{C} \cap \mathrm{SC}(g_{N+1}) = \min(\mathrm{SC}(g_c)), g_c \in \mathrm{C}\} \qquad (6\text{-}2)$$

This strategy is called Minimum Sum of the square of the Correlation (MSC).

In the second strategy, the Maximum value of the square of the Correlation (MC) is calculated as follows:

$$\mathrm{MC}(g_c) \ = \max\,(\mathrm{cor}^2(g_c, g_n)\,), n = 1, 2, \ldots, N. \qquad (6\text{-}3)$$

where, $g_c \in C, g_n \in G_N$.

The selection of $g_{N+1}$ follows the criterion that the MC value is the minimum:

$$\{g_{N+1} \mid g_{N+1} \in \mathrm{C} \cap \mathrm{MC}(g_{N+1}) = \min(\mathrm{MC}(g_c)), g_c \in \mathrm{C}\} \qquad (6\text{-}4)$$

This strategy is called Minimum of Maximum value of the square of the Correlation (MMC).

In the methods mentioned above, a feature is recursively added to the chosen feature set based on supervised learning and the similarity measures. With the use of a classifier in supervised learning, we call the first strategy Classifier-MSC and the second one Classifier-MMC. For example, if the classifier for supervised learning is Naive Bayes Classifier (NBC), we call the two new strategies NBC-MSC and NBC-MMC, respectively.

### 6.2.3 Lagging Prediction Peephole Optimization

Generally, we want to find a subset of features/genes that yields the best classification and prediction performance with the optimal number of genes. The optimization of the feature set in microarray gene expression is highly complicated because of the characterization of the small sample size. Either applying different gene selections to the same training samples or applying the same gene selection to different training samples or applying different learning classifiers to the same training samples will produce different optimization of the feature set. Pochet *et al*. presented a method of determining the optimal number of genes by means of a cross-validation procedure. “In each LOO-CV iteration (number of iterations equals the sample size), one sample is left out of the data, a classification model is trained on the rest of the data and this model is then evaluated on the left out data point” (Pochet *et al.*, 2004). Actually, this procedure by means of LOO-CV utilizes the testing samples in addition to the training samples since the iteration covers all the samples. In the view of my point, the optimization of the number of genes should be just based on the training samples.

The gene selection of RFA is based on supervised learning, with the recursive addition of the next gene; the training classification will increase and finally reach the best classification, and then may maintain it. After that, the training classification may decrease. Normally, all strategies for determining the feature set should be based on the best training classification. If there are multiple best training classifications, just randomly choose one. We call this scheme random strategy under the best training classification. However, in the recursive addition of the features, as training initially reaches the highest accuracy, generally, the training model may not be optimal or robust to the testing samples because of the difference of training samples and the testing samples. In other words, the testing classification may not be the optimal and the best classification model to the testing samples will lag in appearance (see Fig. 1). Based on this consideration and observation, we propose the following algorithm of optimization.

1. Under feature dimension *j,* the training accuracy of the $i^{th}$ experiment is $r(i, j)$. Pick up the feature set $G_k$, corresponding to feature dimension $k$, which has the best training accuracy in the trainings on the feature sets from $G_1$ to $G_D$, corresponding to the feature dimensions from 1 to D. The set of $G_k$ is denoted as HR.

$$\mathrm{HR} = \{G_k \mid 1 \leq \forall\ \mathrm{k} \leq \mathrm{D}, r(i,k) = \max(r(i,j)), 1 \leq \mathrm{j} \leq \mathrm{D}\} \tag{6-5}$$

2. Generally, the best classification model to testing samples will lag in appearance behind the initial best training model. We exclude the elements of HR that correspond to the initial best training. The remaining elements of HR consist of the candidate set HRC for optimization.

3. Each element of HRC is associated with the best training accuracy. We set a peephole on each element and choose the element associated with the best mean value of the training to the whole peephole, described as follows:

a. For each element $G_k \in$ HRC, the peephole on $G_k$ with the length $2l$+1 covers the feature sets $G_{k-l}$, $G_{k-l+1}$,…, $G_k$, ..., $G_{k+l-1}$, $G_{k+l}$, corresponding to the training accuracy $r(i, k-l)$, $r(i, k-l+1)$, …, $r(i, k)$, …,$r(i, k+l-1)$, $r(i, k+l)$. The mean training value of the peephole is denoted as $mp\_r(i,k)$.

$$mp\_r(i,k) = (1/(2l+1)) \sum_{m=k-l}^{m=k+l} r(i,m) \qquad (6\text{-}6)$$

The feature set located on the center of the peephole, which has the best classification of $mp\_r$ is chosen as the optimal one.

b. If there are multiple peepholes with the highest classification $mp\_r$, then we apply random forest to these peepholes and check the mean values of the Out-of-Bag (OOB) error rates [Breiman, 2001; Liaw and Wiener, 2002; Diaz-Uriarte and de Andres, 2006]. The feature sets $G_{k-l}$ ,$G_{k-l+1}$ ,…, $G_{k,}$ ,…, $G_{k+l-1,}$ $G_{k+l}$ correspond to the OOB errors, $oob\_e(i,k-l)$, $oob\_e(i,k-l+1)$,…, $oob\_e(i,k)$, …, $oob\_e(i,k+l-1)$, $oob\_e(i,k+l)$. The mean value of the OOB errors is denoted as $mp\_oob\_e(i,k)$

$$mp\_oob\_e(i,k) = (1/(2l+1)) \sum_{m=k-l}^{m=k+l} oob\_e(i,m) \qquad (6\text{-}7)$$

Pick up the feature set associated with the minimum of $mp\_oob\_e$ as the optimal one.

c. If there are multiple peepholes corresponding to the best $mp\_r$ and minimum $mp\_oob\_e$, then set $l+1 \rightarrow l$, and repeat ‘a’ to ‘c’.

We call this strategy of optimization of RFA as Lagging Prediction Peephole Optimization (LPPO). Fig. 6-1 gives the demonstration of the LPPO on the prostate data set [Singh *et al.*, 2002].

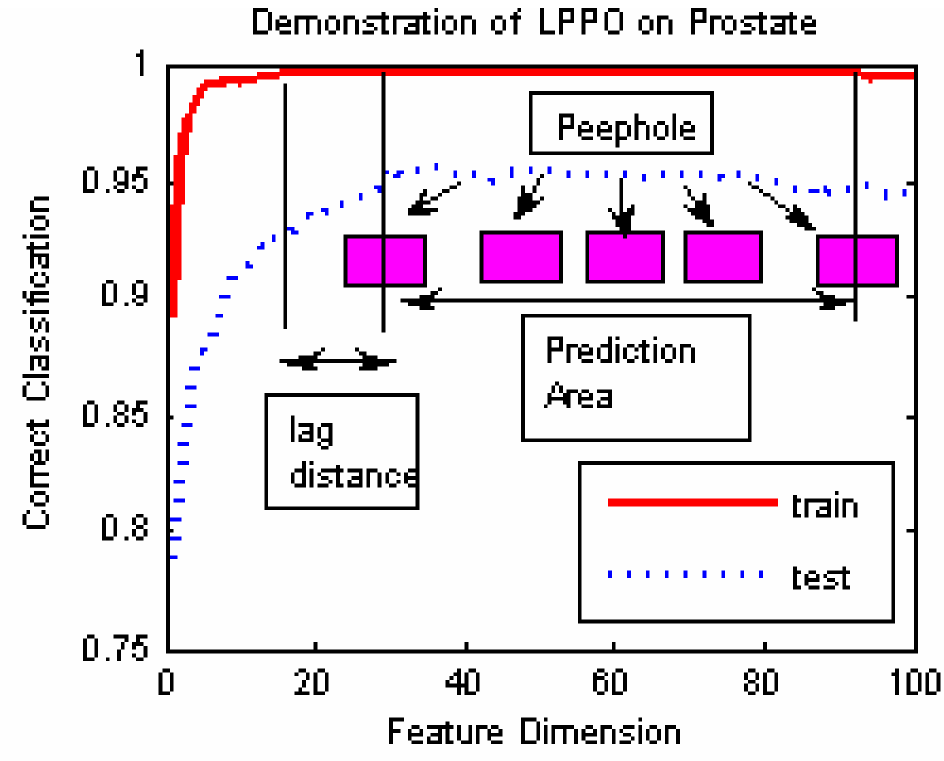


Fig. 6-1 Demonstration of Lagging Prediction Peephole Optimization algorithm on the Prostate data set.

### 6.3 Evaluation of Gene Selection

Under feature dimension *j,* the training accuracy of the $i^{th}$ experiment is $r(i, j)$, and the testing accuracy of the $i^{th}$ experiment is $s(i, j)$, $i$=1, 2, …, $I$; $j$=1, 2, …, $J$; where $I$ is the number of experiments and $J$ is the number of chosen features. The following statistics are measured to evaluate the performance of the gene selections.

(1) The average training accuracy in each feature dimension

The average training accuracy of the experiments under the feature dimension *j*, *r*(*j*), *j*=1, 2, …, *J* is calculated as follows:

$$r(j) = \frac{1}{I} \sum_{i=1}^{I} r(i, j) \tag{6-8}$$

(2) The average testing accuracy in each feature dimension

The average testing accuracy of the experiments under the feature dimension *j*, *s*(*j*), *j*=1, 2, …, *J*, is calculated as follows:

$$s(j) = \frac{1}{I} \sum_{i=1}^{I} s(i, j) \tag{6-9}$$

(3) The average testing accuracy, *ms_hr*(*i*), of the $i^{\text{th}}$ experiment under the condition that the associated/corresponding training accuracy is the highest, which is defined as follows:

$$ms\_hr(i) = \mathrm{mean}(s(i,m)) \mid r(i,m) = \max(r(i,j)), \forall m, j \in \{1,2,..J\} \tag{6-10}$$

Actually, the average testing accuracy *ms_hr*(*i*) is the expected value of the random strategy under the best training classification of the $i^{\text{th}}$ experiment.

(4) The highest testing accuracy, *hs_hr*(*i*), of the $i^{\text{th}}$ experiment under the condition that the associated/corresponding training accuracy is the highest, which is defined as follows:

$$hs\_hr(i) = \max(s(i,m)) \mid r(i,m) = \max(r(i,j)), \forall m, j \in \{1,2,..J\} \tag{6-11}$$

## 6.4 Experiments

### 6.4.1 Data Sets

The following six benchmark microarray gene expression datasets were tested in our experiments. Data sources which are not specified are available at: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

1) The LEUKEMIA data set, which consists of two types of acute leukemia: 48 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloblastic leukemia (AML) samples, over 7129 probes from 6817 human genes [Golub *et al*., 1999].

2) The LYMPHOMA data set, which consists of 58 diffuse large B-cell lymphoma (DLBCL) samples and 19 follicular lymphoma (FL) samples [Shipp *et al*., 2002]. The data file, lymphoma_8_lbc_fscc2_rn.res, and the class label file, lymphoma_8_lbc_fscc2.cls were used in our experiments for identifying DLBCL and FL.

3) The PROSTATE data set contains 52 prostate tumor samples and 50 non-tumor prostate samples [Singh *et al*., 2002].

4) The COLON cancer data set contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of the colons of the same patients. 2000 genes were selected based on the

confidence in the measured expression levels [Alon *et al*., 1999]. The data source is available at http://microarray.princeton.edu/oncology/affydata/index.html.

5) The Central Nervous System (CNS) embryonal tumor data set that was originally studied by [Pomeroy *et al*., 2002]. It contains 60 patient samples. Among them 21 are survivors who are alive after treatment and 39 are failures who succumbed to their diseases. There are 7129 genes.

6) The Breast cancer data set studied by [Van *et al*., 2002]. This data set contains 97 patient samples, 46 patients are relapse who had developed distance metastases within 5 years, and 51 patients are non-relapse who remained healthy for at least 5 years from the distance after their initial diagnosis. This data source is available at: http://www.rii.com/publications/2002/vantveer.htm.

### 6.4.2 Experimental Setup

Our experiments are designed as follows:

1. The data sets are first divided into training samples and testing samples randomly. The ratio of training samples to testing samples is 1:1 in each class.

2. Recursive Feature Additions with Naive Bayes Classifier (NBC) and Nearest Mean Scaled Classifier (NMSC) for gene selection (NBC-MSC, NBC-MMC, NMSC-MSC, and NMSC-MMC) were applied to the training samples for gene selection. Different feature sets of the gene expression data are produced under feature dimensions 1 to 100. We

compared the above proposed methods to several recently developed and published gene selection methods: LOOCSFS, GLGS, SVMRFE, SFFS-LS bound, SFS-LS bound, and also T-TEST.

3. The learning classifiers including NBC, NMSC, SVM, and Random Forest [Breiman, 2001; Liaw and Wiener, 2002] were applied to the testing samples to compare different gene selections.

4. The experiments were performed 20 runs and the average testing accuracies were compared to evaluate performance.

## 6.5 Results

### 6.5.1 Average Training Accuracy

Fig.6-2 lists the average training accuracies on the six data sets with classifiers NMSC, SVM, NBC, and RF. The performances of NBC-MMC, NMSC-MMC, NBC-MSC, and NMSC-MSC are close to one another. Therefore, to clearly demonstrate the other seven gene selections, the average training accuracies of the gene selections NBC-MMC, NMSC-MMC, and NBC-MSC are not presented due to their similar performance in order. Fig.6-2 indicates that on the average with the use of learning classifiers NMSC and NBC, the average training accuracy of NMSC-MSC is the best, followed by GLGS, SVM-RFE, LOOCSFS, SFS-LSbound, SFFS-LSbound, and T-TEST; with the use of learning classifiers SVM and RF, there is no obvious difference in different gene selections.

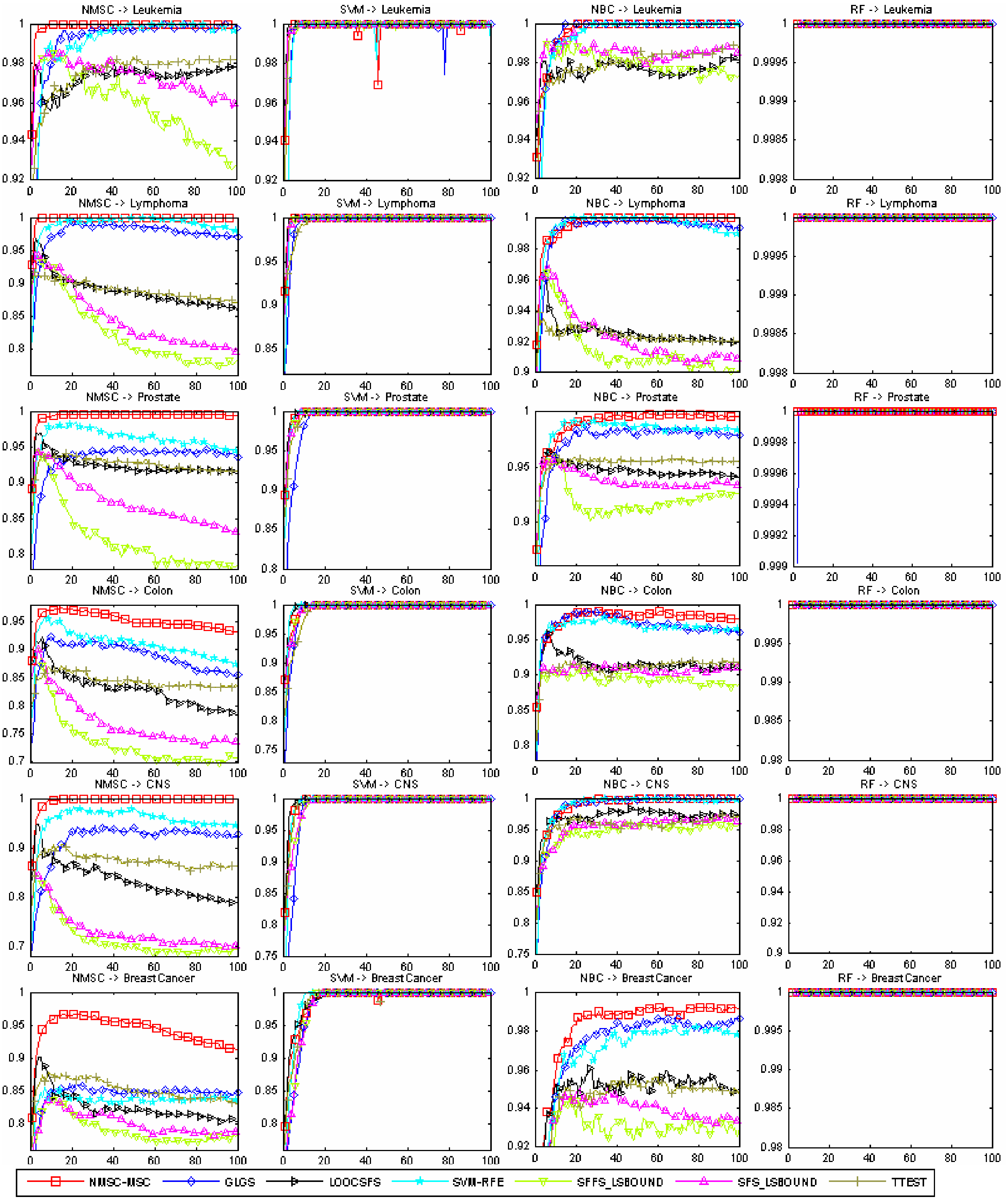


Fig. 6-2 The average training accuracies of different gene selections for six benchmark data sets for four classifiers (NBC, NMSC, SVM, RF). X-axis and y-axis give the feature dimension and testing accuracy values, respectively.

### 6.5.2 Average Testing Accuracy

Fig. 6-3 lists the average testing accuracies of the gene selections with classifiers NMSC, SVM, NBC, and RF. Again, the performances of NBC-MMC, NMSC-MMC, NBC-MSC, and NMSC-MSC are close to one another therefore, the average testing accuracies of the gene selections NBC-MMC, NMSC-MMC, and NBC-MSC are not listed in the figures. Fig 6-3 indicates that, the average testing accuracy of NMSC-MSC is the best, followed by GLGS, LOOCSFS, and SVM-RFE. SFS-LS bound, SFFS-LS bound, and T-TEST didn’t perform well. Fig. 6-3 also manifests that, spanning several data sets and learning classifiers, the performance and stabilization of the gene selection of NMSC-MSC is the best.

### 6.5.3 Testing Accuracy under the Best Training

Table 6-1 provides the mean values and standard errors of the testing accuracies *ms_hr*(*i*), (*i* = 1, 2, …, 20) and the highest testing accuracies *hs_hr*(*i*), (*i*= 1, 2, …, 20) under the highest training classification, defined in (10) and (11), respectively. After applying each classifier to each data set, the highest mean value of the ten gene selections is shaded. In each data set, the highest mean value in the shade is in bold. Table 6-2 lists the statistics of the highest mean value associated with gene selections. Tables 6-1 and 6-2 show that, the best gene selection is NBC-MSC, followed by NMSC-MSC, NMSC-MMC, NBC-MMC, LOOCSFS, GLGS, and SVMRFE. SFFS-LSBOUND and SFS-LSBOUND performed poorly. On the average, T-TEST was the worst.

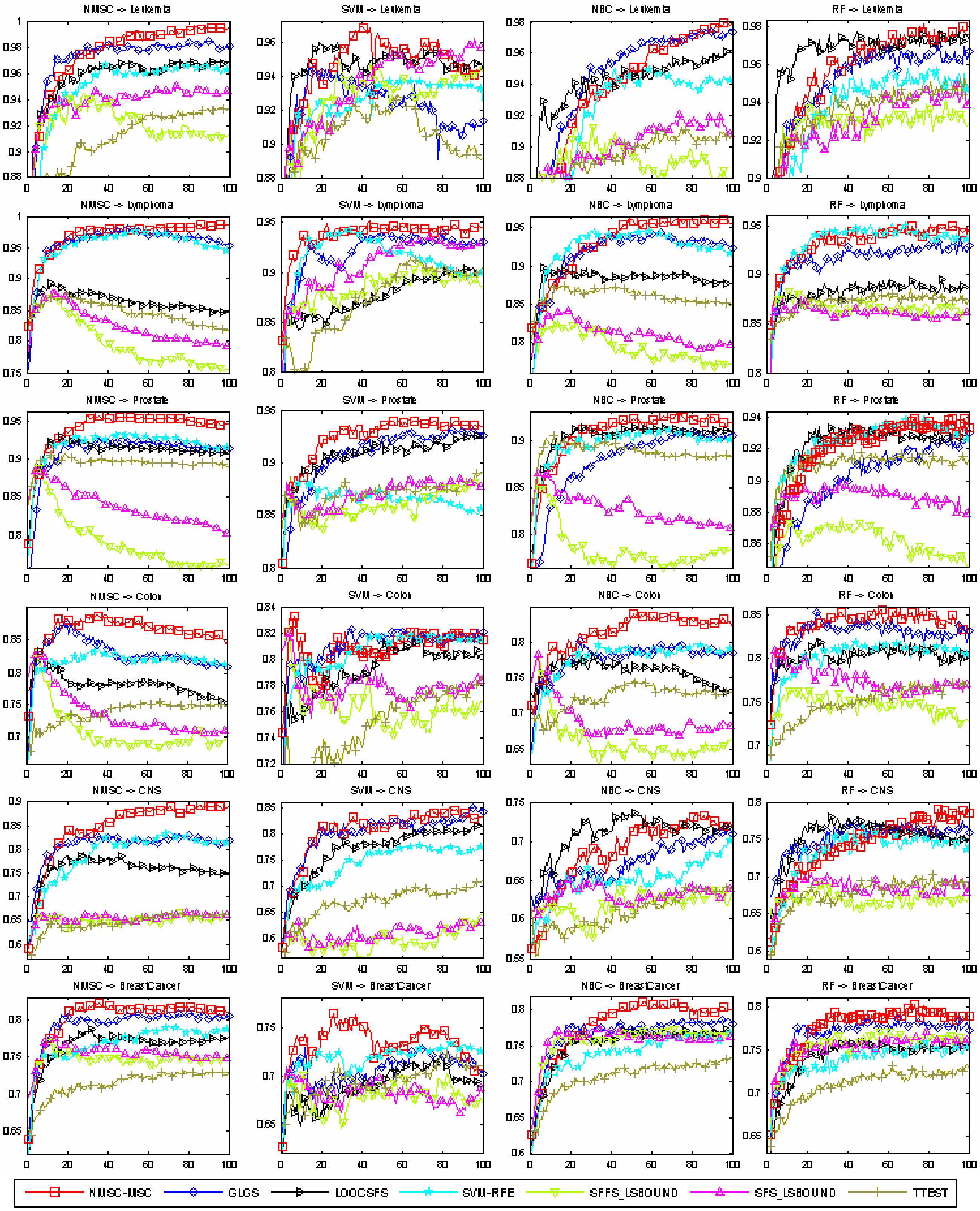


Fig. 6-3 The average testing accuracies of different gene selections for six benchmark data sets for four classifiers (NBC, NMSC, SVM, RF). X-axis and y-axis give the feature dimension and testing accuracy values, respectively.

Table 6-1 Mean values and standard errors of hs_hr and ms_hr. In applying each classifier to each data set, the highest mean value of the ten gene selections is shaded; in each data set, the highest mean value in the shade is in bold.

| DATA SET | GENE SELECTION | MEAN(HS_HR) ± STD(HS_HR), % | | | | MEAN(MS_HR) ± STD(MS_HR), % | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NMSC | SVM | NBC | RF | NMSC | SVM | NBC | RF |
| Leukemia | NBC-MMC | **99.9 ± 0.6** | 99.4 ± 1.2 | 98.3 ± 2.3 | 98.4 ± 1.4 | 98.1 ± 1.4 | 93.4 ± 2.8 | 94.3 ± 2.8 | 95.6 ± 2.3 |
| | NMSC-MMC | **99.9 ± 0.6** | 99.1 ± 1.3 | 98.4 ± 1.9 | 98.6 ± 1.9 | 97.9 ± 1.2 | 93.3 ± 2.8 | 95.2 ± 2.8 | 95.7 ± 3.4 |
| | NBC-MSC | 99.4 ± 1.1 | 99.1 ± 1.3 | 98.9 ± 1.4 | 98.4 ± 1.7 | **98.5 ± 1.6** | 94.9 ± 2.7 | 94.6 ± 2.7 | 96.0 ± 2.5 |
| | NMSC-MSC | 99.7 ± 0.9 | 99.6 ± 1.0 | 98.6 ± 1.7 | 98.7 ± 1.7 | 97.7 ± 1.4 | 94.8 ± 2.5 | 94.6 ± 3.4 | 95.7 ± 3.1 |
| | GLGS | 99.6 ± 1.0 | 98.9 ± 1.7 | 98.6 ± 1.7 | 98.6 ± 1.7 | 97.8 ± 1.7 | 92.5 ± 3.8 | 95.3 ± 1.8 | 95.0 ± 2.5 |
| | LOOCSFS | 97.1 ± 3.3 | 98.0 ± 1.5 | 97.7 ± 1.9 | 99.3 ± 1.2 | 93.9 ± 3.5 | 94.8 ± 3.1 | 94.5 ± 2.7 | 96.7 ± 1.6 |
| | SVMRFE | 98.0 ± 2.0 | 95.4 ± 3.9 | 97.3 ± 2.1 | 98.0 ± 2.0 | 95.7 ± 2.8 | 92.5 ± 5.2 | 92.5 ± 3.0 | 93.4 ± 1.9 |
| | SFFS-LSBOUND | 97.1 ± 2.5 | 97.4 ± 3.8 | 96.3 ± 4.1 | 97.1 ± 2.8 | 93.8 ± 4.3 | 92.9 ± 3.8 | 90.2 ± 5.8 | 92.6 ± 4.1 |
| | SFS-LSBOUND | 97.1 ± 2.8 | 97.0 ± 3.0 | 96.4 ± 3.6 | 97.3 ± 3.0 | 94.6 ± 3.5 | 93.6 ± 3.8 | 91.2 ± 5.0 | 93.0 ± 5.1 |
| | T-TEST | 94.8 ± 3.5 | 95.4 ± 4.5 | 93.3 ± 6.9 | 96.8 ± 2.9 | 92.2 ± 3.9 | 90.7 ± 4.8 | 90.1 ± 6.5 | 93.5 ± 3.6 |
| Lymphoma | NBC-MMC | 98.1 ± 2.6 | 99.0 ± 1.3 | 97.3 ± 2.6 | 96.4 ± 2.8 | 96.2 ± 4.3 | 93.8 ± 2.8 | 91.7 ± 3.9 | 91.6 ± 3.7 |
| | NMSC-MMC | 99.2 ± 1.2 | 98.8 ± 1.6 | 97.9 ± 2.6 | 96.5 ± 3.7 | 96.9 ± 1.9 | 93.0 ± 2.8 | 93.1 ± 3.3 | 92.3 ± 4.0 |
| | NBC-MSC | 99.4 ± 1.1 | 98.4 ± 1.8 | 97.9 ± 2.6 | 96.8 ± 3.3 | **97.5 ± 1.9** | 93.1 ± 3.5 | 92.7 ± 3.5 | 92.6 ± 4.1 |
| | NMSC-MSC | **99.5 ± 1.1** | 98.8 ± 1.6 | 98.1 ± 2.0 | 97.0 ± 3.6 | 97.2 ± 1.9 | 93.9 ± 3.0 | 93.9 ± 3.1 | 93.4 ± 3.9 |
| | GLGS | 98.6 ± 1.8 | 98.2 ± 1.9 | 97.0 ± 2.6 | 96.9 ± 2.3 | 96.5 ± 2.1 | 92.5 ± 3.8 | 92.3 ± 3.6 | 91.7 ± 2.9 |
| | LOOCSFS | 87.0 ± 7.2 | 93.0 ± 5.3 | 87.3 ± 5.1 | 92.9 ± 4.8 | 85.8 ± 6.8 | 87.8 ± 5.4 | 85.1 ± 4.5 | 88.2 ± 4.3 |
| | SVMRFE | 99.2 ± 1.5 | 96.5 ± 3.9 | 97.2 ± 3.4 | 96.6 ± 3.1 | 96.5 ± 2.0 | 91.8 ± 4.3 | 93.1 ± 4.0 | 93.3 ± 4.0 |
| | SFFS-LSBOUND | 88.7 ± 6.1 | 95.1 ± 3.3 | 84.0 ± 4.9 | 92.2 ± 4.7 | 87.0 ± 5.7 | 88.2 ± 4.9 | 80.6 ± 3.9 | 86.8 ± 4.8 |
| | SFS-LSBOUND | 87.7 ± 6.1 | 96.1 ± 3.5 | 86.1 ± 3.5 | 91.8 ± 4.2 | 86.4 ± 5.6 | 91.1 ± 3.7 | 82.7 ± 3.4 | 86.1 ± 4.8 |
| | T-TEST | 86.0 ± 5.7 | 94.4 ± 3.0 | 86.5 ± 7.0 | 91.7 ± 5.2 | 84.3 ± 5.8 | 87.7 ± 3.3 | 83.9 ± 6.1 | 87.2 ± 4.5 |
| Prostate | NBC-MMC | 96.3 ± 2.4 | 95.8 ± 2.5 | 94.8 ± 2.6 | 96.5 ± 2.0 | 94.2 ± 2.8 | 91.6 ± 2.3 | 90.4 ± 2.7 | 92.1 ± 2.2 |
| | NMSC-MMC | 95.6 ± 2.3 | 95.9 ± 2.5 | 93.7 ± 2.8 | 95.3 ± 2.3 | 92.7 ± 2.3 | 91.4 ± 2.8 | 90.7 ± 3.1 | 91.3 ± 2.3 |
| | NBC-MSC | 96.4 ± 2.0 | 96.6 ± 1.9 | 95.2 ± 2.1 | 96.5 ± 1.9 | **94.6 ± 2.3** | 92.5 ± 2.3 | 91.0 ± 2.3 | 92.5 ± 2.2 |
| | NMSC-MSC | **96.9 ± 2.3** | 96.7 ± 1.7 | 94.5 ± 2.0 | 95.8 ± 1.8 | 94.5 ± 2.4 | 92.8 ± 1.9 | 91.8 ± 2.5 | 92.0 ± 1.9 |
| | GLGS | 93.6 ± 3.0 | 96.1 ± 2.2 | 90.4 ± 3.9 | 94.7 ± 2.0 | 91.5 ± 2.7 | 91.7 ± 2.6 | 87.5 ± 3.4 | 90.0 ± 2.5 |
| | LOOCSFS | 88.4 ± 5.2 | 94.9 ± 2.9 | 90.7 ± 5.3 | 95.2 ± 2.6 | 87.0 ± 4.7 | 91.1 ± 3.4 | 88.0 ± 4.5 | 92.3 ± 2.3 |
| | SVMRFE | 94.1 ± 3.4 | 92.3 ± 2.7 | 92.8 ± 4.3 | 95.7 ± 2.6 | 92.4 ± 3.3 | 86.7 ± 3.5 | 90.0 ± 4.0 | 92.5 ± 2.8 |
| | SFFS-LSBOUND | 90.4 ± 3.2 | 93.4 ± 2.8 | 86.2 ± 5.8 | 90.2 ± 3.2 | 88.9 ± 3.1 | 86.0 ± 3.2 | 84.4 ± 5.1 | 86.1 ± 4.0 |
| | SFS-LSBOUND | 89.7 ± 4.9 | 92.7 ± 4.0 | 87.3 ± 5.4 | 92.4 ± 3.5 | 88.3 ± 5.1 | 87.2 ± 5.0 | 85.1 ± 5.4 | 89.0 ± 3.9 |
| | T-TEST | 91.4 ± 4.1 | 92.5 ± 2.1 | 91.7 ± 2.8 | 94.0 ± 3.0 | 89.7 ± 3.7 | 87.1 ± 3.2 | 89.0 ± 4.3 | 91.0 ± 3.1 |
| Colon | NBC-MMC | 88.7 ± 5.5 | 87.7 ± 5.2 | 86.5 ± 4.0 | 89.7 ± 4.9 | 84.5 ± 5.2 | 80.9 ± 6.0 | 78.2 ± 4.9 | 82.5 ± 5.5 |
| | NMSC-MMC | **91.1 ± 5.0** | 87.7 ± 3.9 | 87.4 ± 5.3 | 90.0 ± 4.0 | 84.9 ± 7.1 | 81.3 ± 5.5 | 80.8 ± 5.9 | 83.3 ± 5.4 |
| | NBC-MSC | 89.4 ± 4.3 | 86.9 ± 4.6 | 88.7 ± 6.0 | 90.0 ± 4.0 | **86.0 ± 5.2** | 80.3 ± 5.6 | 82.1 ± 4.8 | 84.4 ± 4.7 |
| | NMSC-MSC | 91.0 ± 5.3 | 87.6 ± 4.7 | 88.1 ± 3.3 | 90.0 ± 4.4 | **86.0 ± 5.4** | 80.9 ± 5.5 | 82.6 ± 4.0 | 83.9 ± 4.5 |
| | GLGS | 87.3 ± 6.2 | 87.3 ± 4.6 | 85.2 ± 4.8 | 90.5 ± 4.3 | 83.7 ± 6.6 | 81.2 ± 5.5 | 77.6 ± 5.8 | 83.0 ± 4.5 |
| | LOOCSFS | 85.0 ± 5.3 | 86.3 ± 3.9 | 81.6 ± 5.8 | 86.8 ± 5.3 | 82.2 ± 4.6 | 79.3 ± 5.3 | 76.7 ± 6.9 | 80.3 ± 5.3 |
| | SVMRFE | 86.0 ± 6.7 | 86.8 ± 4.8 | 82.1 ± 7.4 | 86.3 ± 5.5 | 81.8 ± 7.2 | 80.7 ± 4.7 | 77.7 ± 7.5 | 80.3 ± 6.0 |
| | SFFS-LSBOUND | 85.0 ± 4.8 | 87.1 ± 4.4 | 72.7 ± 7.0 | 82.6 ± 6.0 | 82.4 ± 4.4 | 76.2 ± 6.3 | 69.5 ± 8.3 | 74.6 ± 6.8 |
| | SFS-LSBOUND | 85.3 ± 4.6 | 85.8 ± 5.3 | 76.8 ± 7.1 | 86.0 ± 4.1 | 83.3 ± 4.7 | 77.7 ± 6.4 | 72.5 ± 6.2 | 77.6 ± 4.5 |
| | T-TEST | 77.4 ± 10.4 | 85.5 ± 4.0 | 76.3 ± 8.3 | 81.5 ± 7.2 | 74.9 ± 10.8 | 75.3 ± 5.7 | 72.8 ± 8.2 | 75.1 ± 7.8 |
| CNS | NBC-MMC | 91.8 ± 6.1 | 92.9 ± 3.6 | 77.8 ± 5.2 | 85.7 ± 4.0 | 86.7 ± 6.0 | 82.4 ± 4.7 | 67.3 ± 4.1 | 76.3 ± 4.0 |
| | NMSC-MMC | 90.0 ± 6.4 | 92.2 ± 5.7 | 78.0 ± 5.3 | 82.7 ± 5.2 | 82.8 ± 6.8 | 82.1 ± 5.6 | 67.5 ± 5.5 | 73.5 ± 4.9 |
| | NBC-MSC | **94.0 ± 4.6** | 92.0 ± 4.4 | 81.1 ± 4.1 | 85.5 ± 4.9 | **88.4 ± 5.2** | 82.6 ± 5.5 | 70.2 ± 3.7 | 75.9 ± 5.3 |
| | NMSC-MSC | 92.8 ± 4.0 | 91.6 ± 4.9 | 81.3 ± 6.1 | 84.9 ± 4.1 | 85.6 ± 4.3 | 81.4 ± 6.2 | 70.0 ± 4.5 | 74.4 ± 4.2 |
| | GLGS | 84.7 ± 3.3 | 91.1 ± 5.4 | 78.8 ± 5.5 | 84.2 ± 5.0 | 82.4 ± 3.6 | 81.3 ± 4.8 | 67.9 ± 4.5 | 75.3 ± 4.3 |
| | LOOCSFS | 71.3 ± 9.8 | 85.0 ± 5.9 | 79.1 ± 7.7 | 83.2 ± 4.4 | 69.3 ± 8.0 | 77.6 ± 4.5 | 71.8 ± 6.2 | 75.3 ± 5.1 |
| | SVMRFE | 83.2 ± 8.9 | 85.1 ± 8.4 | 77.1 ± 6.8 | 83.5 ± 4.3 | 77.0 ± 8.0 | 75.0 ± 8.8 | 65.7 ± 7.2 | 73.3 ± 4.9 |
| | SFFS-LSBOUND | 68.1 ± 6.7 | 71.9 ± 7.1 | 67.6 ± 7.7 | 76.2 ± 4.5 | 65.3 ± 6.3 | 59.4 ± 7.5 | 61.3 ± 6.1 | 66.9 ± 4.8 |
| | SFS-LSBOUND | 67.8 ± 6.2 | 72.4 ± 4.9 | 69.8 ± 8.2 | 76.2 ± 5.0 | 65.7 ± 5.4 | 60.7 ± 5.1 | 63.7 ± 7.2 | 68.4 ± 4.5 |
| | T-TEST | 67.5 ± 8.8 | 77.4 ± 6.4 | 67.0 ± 7.1 | 75.5 ± 5.9 | 63.4 ± 7.6 | 67.3 ± 5.8 | 60.9 ± 6.8 | 67.8 ± 4.9 |
| Breast | NBC-MMC | 82.5 ± 6.0 | 82.9 ± 3.5 | 84.1 ± 3.0 | 84.1 ± 3.6 | 81.3 ± 5.7 | 73.2 ± 3.8 | 78.4 ± 3.4 | 78.4 ± 3.8 |
| | NMSC-MMC | 83.9 ± 4.6 | 82.0 ± 3.3 | 82.4 ± 4.3 | 83.7 ± 4.7 | 80.4 ± 4.0 | 72.0 ± 3.8 | 78.4 ± 4.3 | 77.0 ± 4.3 |
| | NBC-MSC | 83.4 ± 5.8 | 83.5 ± 3.8 | 85.8 ± 3.1 | **85.9 ± 4.7** | **81.5 ± 5.3** | 74.9 ± 3.3 | 79.1 ± 3.0 | 79.4 ± 4.1 |
| | NMSC-MSC | 82.8 ± 4.4 | 82.4 ± 3.8 | 84.1 ± 4.0 | 83.9 ± 4.0 | 79.6 ± 4.0 | 73.7 ± 3.9 | 79.2 ± 3.8 | 77.7 ± 4.0 |
| | GLGS | 80.8 ± 3.7 | 79.3 ± 4.5 | 81.4 ± 4.1 | 83.7 ± 4.6 | 79.2 ± 3.9 | 70.7 ± 4.6 | 77.8 ± 3.7 | 77.0 ± 4.2 |
| | LOOCSFS | 71.7 ± 6.5 | 77.3 ± 5.2 | 78.0 ± 5.8 | 80.3 ± 3.8 | 70.4 ± 6.5 | 69.2 ± 4.7 | 74.7 ± 5.1 | 74.3 ± 4.2 |
| | SVMRFE | 74.3 ± 7.1 | 78.3 ± 5.2 | 77.2 ± 5.3 | 80.4 ± 4.1 | 73.2 ± 6.6 | 72.1 ± 5.8 | 73.9 ± 4.5 | 73.9 ± 3.7 |
| | SFFS-LSBOUND | 76.2 ± 5.2 | 78.9 ± 2.8 | 76.9 ± 7.3 | 81.5 ± 5.3 | 75.0 ± 5.3 | 67.8 ± 3.3 | 75.2 ± 6.8 | 75.6 ± 4.9 |
| | SFS-LSBOUND | 77.5 ± 5.6 | 78.9 ± 4.2 | 79.8 ± 5.2 | 81.3 ± 5.2 | 75.8 ± 5.5 | 68.0 ± 4.7 | 76.9 ± 6.3 | 75.4 ± 5.2 |
| | T-TEST | 71.1 ± 5.3 | 77.6 ± 5.2 | 72.6 ± 6.3 | 76.3 ± 5.7 | 69.3 ± 5.3 | 69.9 ± 3.6 | 70.5 ± 5.8 | 71.1 ± 5.8 |

Table 6-2 Numbers of occurrences of the highest mean values in Table 6-1

| Gene Selection | # of shade | | # of bold | |
|---|---|---|---|---|
| | HS HR | MS HR | HS HR | MS HR |
| NBC-MMC | 6 | 1 | 1 | 0 |
| NMSC-MMC | 4 | 1 | 2 | 0 |
| NBC-MSC | 8 | 12 | 2 | 6 |
| NMSC-MSC | 7 | 8 | 2 | 1 |
| GLGS | 1 | 1 | 0 | 0 |
| LOOCSFS | 1 | 2 | 0 | 0 |
| SVMRFE | 0 | 1 | 0 | 0 |
| SFFS-LSBOUND | 0 | 0 | 0 | 0 |
| SFS-LSBOUND | 0 | 0 | 0 | 0 |
| T-TEST | 0 | 0 | 0 | 0 |
| Total | 27 | 26 | 7 | 7 |

Table 6-3 Comparison of LPPO and Random Strategy

| Data Set | Gene Selection | MEAN(S_LPPO - MS_HR) , % | | | |
|---|---|---|---|---|---|
| | | NMSC | SVM | NBC | RF |
| Leukemia | NBC-MMC | 0.8 | -0.1 | 2.3 | 1.4 |
| | NMSC-MMC | 1.0 | 0.9 | 1.8 | 1.6 |
| | NBC-MSC | -0.2 | 0.3 | 1.9 | 1.1 |
| | NMSC-MSC | 1.6 | 0.7 | 2.5 | 1.3 |
| Lymphoma | NBC-MMC | 0.6 | 0.1 | -1.0 | 0.4 |
| | NMSC-MMC | 1.3 | -0.4 | 1.4 | 1.2 |
| | NBC-MSC | 0.4 | 1.2 | 1.5 | 1.4 |
| | NMSC-MSC | 0.9 | 0.1 | 1.6 | 0.6 |
| Prostate | NBC-MMC | 0.2 | 0.1 | 0.0 | 0.5 |
| | NMSC-MMC | 0.9 | 0.4 | 0.9 | 1.1 |
| | NBC-MSC | 0.3 | 0.7 | 0.6 | 1.8 |
| | NMSC-MSC | 0.4 | 0.8 | 0.2 | 1.0 |
| Colon | NBC-MMC | 0.3 | 0.2 | -1.1 | 0.4 |
| | NMSC-MMC | 0.6 | 0.0 | 0.1 | 0.3 |
| | NBC-MSC | -0.2 | -0.5 | -2.6 | -1.3 |
| | NMSC-MSC | 0.9 | 0.3 | -2.2 | -0.5 |
| CNS | NBC-MMC | 2.1 | 1.8 | 2.2 | 3.1 |
| | NMSC-MMC | 0.8 | 1.0 | 0.4 | 1.6 |
| | NBC-MSC | 1.2 | 0.0 | 0.6 | 0.6 |
| | NMSC-MSC | 1.9 | 2.2 | 2.4 | 1.3 |
| Breast Cancer | NBC-MMC | 0.2 | 1.3 | 0.5 | 1.5 |
| | NMSC-MMC | 0.6 | 3.2 | -1.2 | 0.9 |
| | NBC-MSC | 0.0 | 1.7 | -1.6 | -0.6 |
| | NMSC-MSC | 1.7 | 1.3 | -1.1 | 1.0 |
| **Average** | | **0.8** | **0.7** | **0.4** | **0.9** |

### 6.5.4 Comparison of LPPO and Random Strategy

Table 6-3 lists the mean values of the differences between the testing values (denoted as S_LPPO) by applying NMSC, SVM, NBC, and RF to LPPO and ms_hr. The table shows that, on the average, LPPO is superior to the random strategy under the best training acuuracies. In summary, spanning the six benchmark data sets, in comparison with ms_hr, LPPO improves the testing accuracy by an average of 0.8% for NMSC, 0.7% for SVM, 0.4% for NBC, and 0.9% for RF.

### 6.5.5 Comparison of LPPO and varSelRF

Fig. 6-4 shows the boxplots of the testing values of the feature sets LPPO with RFA and varSelRF with RF. The gene selections are NBC-MMC, NMSC-MMC, NBC-MSC, NMSC-MSC, and varSelRF from left to right in each subfigure. Fig. 6-4 shows that the testing accuracy values by applying RF to the feature set of LPPO on RFA are higher than the values by applying RF to the feature set from the gene selection of varSelRF.

## 6.6 CONCLUSION

This chapter presents a new gene selection method: Recursive Feature Addition for improving classifications of microarray gene expression data. This method takes advantage of the highest training accuracy and adds the subsequent gene recursively based on the similarity measures between the chosen genes and the candidates in order to minimize the redundancy of the genes within the selected subset of genes. In order to have a fair comparison across all methods, we addressed the issue of optimizing the number of genes for each of the methods. We proposed the Lagging Prediction Peephole

Optimization algorithm for optimizing the number of genes and to choose the final feature/gene set. We compared RFA to other gene selection methods using six popular benchmark datasets. Results show that, RFA outperforms the other recently developed methods with the use of different classifiers. Results also show that, on the average, the testing accuracy with the feature set chosen by LPPO is superior to the random strategy under the best training accuracies. Regarding the classification accuracy, LPPO also outperforms the popular gene selection method varSelRF.

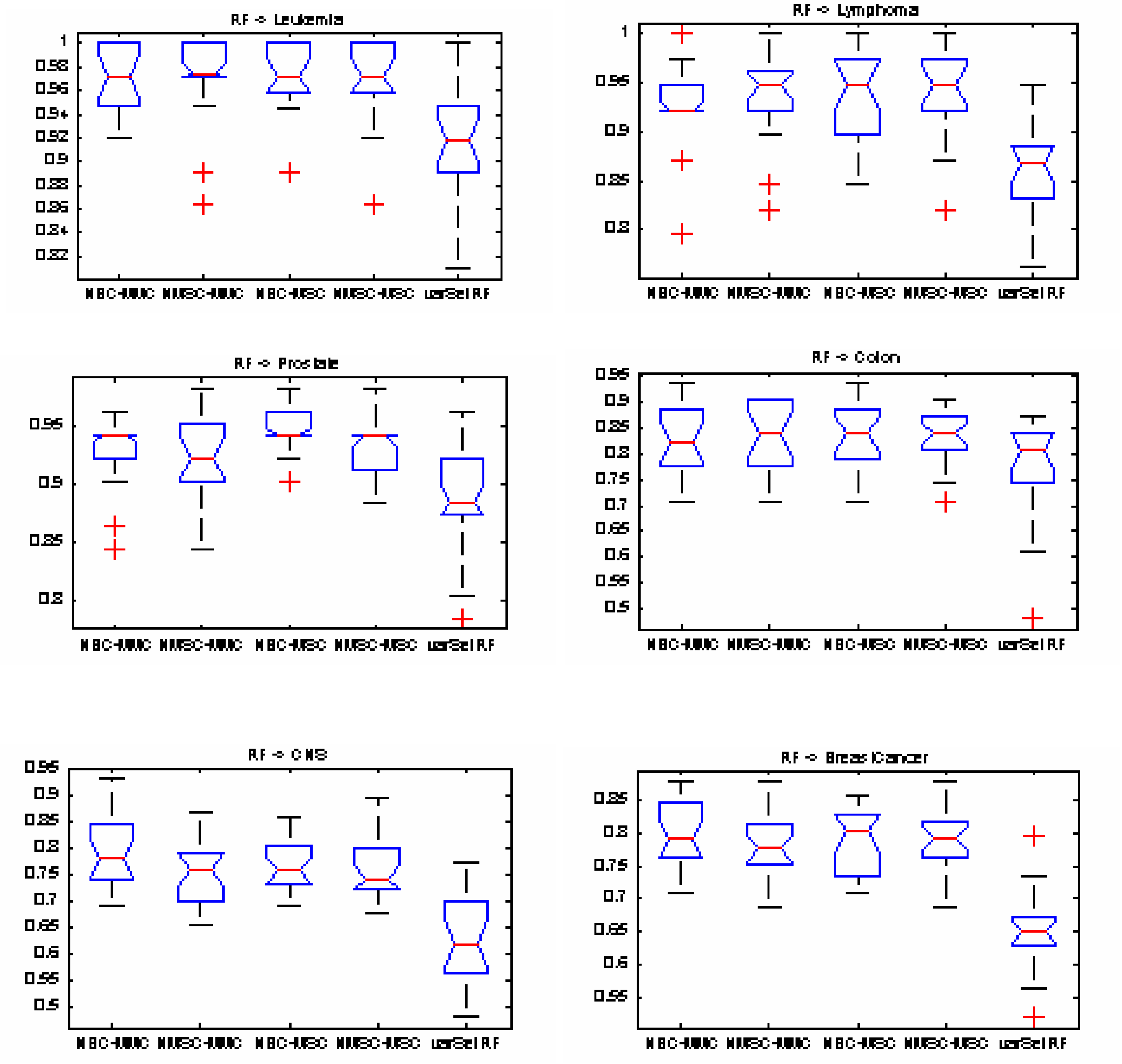


Fig. 6-4 Boxplots of testing accuracies of the LPPO with RFA VS varSelRF for six data sets. Random Forest is the testing classifier.

# CHAPTER 7: TAGGING SNP SELECTION FOR GENOME-WIDE DISEASE CLASSIFICATION

## 7.1 Introduction

SNPs promise to significantly advance our ability to understand and treat human disease. Comprehensive evaluation of common genetic variations through association of SNP structure with common complex diseases in the genome-wide scale is currently a hot area in human genome research. However, due to the tremendous number of candidate SNPs, there are a clear need to expedite genotyping by selecting and considering only a subset of all SNPs. This process is known as ***tagging SNP selection***. Exploiting information redundancy due to associations between single nucleotide polymorphism (SNP) markers potentially reduces the efforts in terms of time and cost for these studies. One of the fundamental questions in SNP-disease association study is how many SNPs is enough to provide good prediction performance of disease status. This chapter presents a new feature selection method named Supervised Recursive Feature Addition (SRFA). This method combines supervised learning and statistical measures for the chosen candidate features/SNPs to deal with the redundancy information so that it can improve the classification in association studies. Additionally, this chapter also describes a Support Vector based lowest weight and lowest correlation Recursive Feature Addition (SVFRA) scheme in SNP-diseases association analysis. We implemented the proposed SRFA with different statistical learning classifiers for both SNP selections and disease classifications, and then applied them to two complex disease data sets. Results show that on the average, designed SRFA outperforms the well-known method of Support Vector

Machine Recursive Feature Elimination and logic regression based SNP selections for disease classification in genetic association study.

## 7.2 Related Work

Correlating variations in DNA sequence with phenotypic differences has been one of the grand challenges in biomedical research. Substantial efforts have been made to obtain all common genetic variations in humans, including single nucleotide polymorphisms (SNPs), deletions and insertions [Brookes, 1999]. The HapMap Project has collected genotypes of millions of SNPs from populations with ancestry from Africa, Asia and Europe and makes this information freely available in the public domain [The International HapMap Consortium, 2003, 2004, 2005]. Yet, one cannot perform a whole genome-wide association study directly based on the genotypes or allele frequencies of individual markers due to the relative low power of each SNP and the huge number of total SNPs. While millions of SNPs have been identified, with an estimated two common missense variants per gene, there is a great need, conceptually as well as computationally, to develop advanced robust algorithms and analytical methods for characterizing genetic variations that are non-redundant and identify the target SNPs that are most likely to affect the phenotypes and ultimately contribute to disease development.

Exploiting information redundancy due to associations between SNP markers potentially reduces the efforts in terms of time and cost for genetic association studies [Risch, 2000]. However, the efficacy of searching for optimal set of SNPs has not been as successful as expected in theory. One primary cause is the high dimensionality with highly correlated features/SNPs that can hinder the power of the identification of small to moderate genetic effects in complex diseases. The need to incorporate covariates of other environmental risk factors as effect modifiers or confounders further worsens “the curse of dimensionality problem” in mapping genes for complex diseases [Cardon and Bell, 2001]. One of the fundamental questions for searching for set of SNPs in genetic association study is how many SNPs is enough to provide good prediction performance of disease status.

Therefore, feature selection for massive genomic data in high dimension has become a main task to be tackled with statistical and computational efforts recently. Specifically, in genome-wide disease association studies, various models and algorithms have been proposed for selecting a subset of SNPs [Hampe et al., 2003; Sebastiani et al., 2003; Stram et al., 2003; Carlson *et al*., 2004; Halldorsson et al., 2004; Lin and Altman, 2004; Goplakrishnan and Qin, 2006]. Linkage Disequilibrium based methods for selecting a maximally informative set of SNPs for association analyses has been developed first [Cores and Vapnik, 1995; Vapnik, 1995; Vapnik 1998; Witte and Fijal, 2001; Tan *et al*., 2005]. Zhang and Jin introduced a tagSNPs criterion based on pair-wise Linkage Disequilibrium (LD) and haplotype $r^2$ measure for case control association studies [Zhan and Jin, 2003]. [Anderson and Novermbre, 2003] and [Mannila *et al*., 2003] proposed finding haplotype block boundaries using minimum description length. The method presented by [Beckmann *et al*., 2005] reflects the flexibility of Mantel statistics using haplotype sharing to correlate temporal and spatial distributions of cancer in a generalized regression approach for SNP selections and disease mapping purposes. The tagSNPs for unphased genotypes is designed based on multiple linear regressions [He and Zelikovsky, 2006]. Other test statistic approaches such as scan statistic by [Levin *et al*., 2005]; score statistic by [Schaid *et al*., 2002], weighted-average statistic [Song and Elston, 2006] for disease mapping in case-control studies were proposed for SNP selection in genetic association studies.

Recently, Schwender and Ickstadt demonstrated logic regression [Kooperberg *et al*., 2001] based identification of SNP interactions for the disease status in case-control study and proposed two measures for quantifying the importance of feature interactions for classification. In comparison with some well-known classification methods of CART [Breiman *et al*., 1984], Random Forests [Breiman, 2001] and other regression procedures [Witte and Fijal, 2001], logic regression has shown a good classification performance when applied to SNP data [Schwender and Ickstadt, 2006].

In this chapter, a new feature selection method named Supervised Recursive Feature Addition (SRFA) is presented. This method combines supervised learning and statistical measures for the chosen candidate features/SNPs in order to deal with the redundancy information so that it can improve the classification and prediction performance. We implemented our SRFA with different statistical learning classifiers for both SNP selections and disease classifications and compared their performances to popular classification models, such as conditional logistic regression, Logic regression, and Support Vector Machine Recursive Feature Elimination (SVMRFE). Additionally, we propose a support vector based lowest weight and lowest correlation feature selection scheme for SNP-diseases association analysis. We applied these proposed approaches to two complex SNP-disease data sets: Myocardial Infarction Case & Control (MICC) data set and a subset of The North American Rheumatoid Arthritis Consortium (NARAC) data to evaluate and to demonstrate our proposed SRFA with different supervised learning classifiers for both SNP selections and disease classifications.

## 7.3 Supervised Tagging SNP Selection

### 7.3.1 Supervised Recursive Feature Addition Algorithm for SNP Selection

SRFA combines supervised learning and statistical similarity measures between the chosen features and the candidates and is presented as follows:

Step 1: Each individual feature is ranked from the highest classification accuracy to the lowest classification accuracy with the use of a supervised learning classifier.

Step 2: The feature with the highest classification accuracy is chosen as the first feature. If multiple features achieve the same highest classification accuracy, the one with the lowest *p*-value measured by score test-statistics is chosen as the first element. At this point the chosen feature set, $G_1$, consists of the first feature, $g_1$, which corresponds to feature dimension one.

Step 3: The $N+1$ – dimensional feature set, $G_{N+1} = \{g_1, g_2, \ldots, g_N, g_{N+1}\}$ is produced by adding $g_{N+1}$ to the previous $N$-dimensional feature set, $G_N = \{g_1, g_2, \ldots, g_N\}$. $g_{N+1}$ is chosen as follows:

Temporarily add each feature $g_i$ ($i \neq 1, 2, \ldots, N$) outside of $G_N$ to $G_N$: The classification accuracies of each feature set $G_N + \{g_i\}$ is recorded, the $g_c$ with the highest classification accuracy is marked and put into the set of candidates: $C$. Generally, the set of candidates consists of many features, but only one feature will be selected to be included in the feature set next as $g_{N+1}$. We choose the $(N+1)^{\text{th}}$ feature: $g_{N+1}$ from candidate set $C$ according to statistical similarity between the chosen features and candidates. We call this step Candidate Feature Addition. The goal is to obtain a most informative and least redundant feature set. The statistical similarity measure is based on the Spearman Correlation Coefficient (for categorical features/SNPs) between the chosen feature $g_n$ ($g_n \in G_N$, $n = 1, 2, \ldots, N$) and the candidate $g_c$ ($g_c \in C$, $c = 1, 2 \ldots m$; $m$ is the number of elements in $C$). Spearman's rank correlation coefficient, often denoted by the Greek letter ρ (rho), is a non-parametric measure of correlation – that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike the Pearson product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level.

In principle, ρ is simply a special case of the Pearson product-moment coefficient in which the data are converted to ranks before calculating the coefficient. In practice, however, a simpler procedure is normally used to calculate ρ. The raw scores are converted to ranks, and the differences $D$ between the ranks of each observation on the two variables are calculated. ρ is then given by:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad (7\text{-}1)$$

where, $D$ = the difference between the ranks of corresponding values of $X$ and $Y$, and $N$ = the number of pairs of values.

The Sum of the square of the Correlation (SC) is calculated to measure the similarity and is defined as follows:

$$\text{SC}(g_c) = \sum_{n=1}^{N} \rho^2(g_c, g_n), \; n = 1, 2 \ldots N \qquad (7\text{-}2)$$

where $g_c \in C, g_n \in G_N$.

The selection of $g_{N+1}$ follows the qualification that the SC value in (7-2) is the minimum:

$$\{g_{N+1} \mid g_{N+1} \in C \cap \text{SC}(g_{N+1}) = \min(\text{SC}(g_c)), g_c \in C\} \qquad (7\text{-}3)$$

This strategy is called Minimum SC (MSC).

Step 4: A feature is recursively added to the chosen feature set from steps 1-3 with supervised learning and the similarity measures until classification accuracy stops to increase.

Our SRFA based MSC is denoted as classifier-MSC, for example, if the classifier is Naive Bayes Classifier (NBC), we call the feature selection NBC-MSC. SRFA here can not only provide us the feature selection procedure but also it could be directly used for further classification and prediction purposes by using learning classifiers that may differ from feature selection classifiers.

### 7.3.2 Support Vector Based Recursive Feature Addition Algorithms

Support Vector Machines (SVMs) [Cores and Vapnik, 1995; Vapnik, 1995] have been widely applied to pattern classification problems and non-linear regressions. The basic idea of the SVM algorithm is to find an optimal hyper-plane that can maximize the margin between two groups. The vectors that are closest to the optimal hyper-plane are called support vectors. [Guyon *et al*., 2002] proposed a feature selection, called Support Vector Machine Recursive Feature Elimination (SVMRFE). Based on the SVMRFE and our SRFA discussed earlier, we propose a Support Vector based lowest weight (or maximum margin width) and lowest correlation feature addition scheme, called Support Vector based Recursive Feature Addition (SVRFA) described as follows:

1. Train an SVM on each individual feature in the data set given an SVM with weight vector $\vec{w} = \sum_k \alpha_k y_k \overrightarrow{x_k}$

2. Rank features according to criterion $c$ for feature $i$: $c_i = (w_i)^2$. The features corresponding to the lowest $c$ are picked up as candidates. The candidate with the highest statistical significance is the first element of the feature set. At this point the chosen

feature set, $G_1$, consists of the first feature, $g_1$, which corresponds to feature dimension one.

3. The $(N+1)^{st}$ dimensional feature set, $G_{N+1} = \{g_1, g_2, \ldots, g_N, g_{N+1}\}$ is produced by adding $g_{N+1}$ to the $N$ dimensional feature set, $G_N = \{g_1, g_2, \ldots, g_N\}$. The choice of $g_{N+1}$ is described as follows:

Temporarily add each feature $g_i$ ($i \neq 1, 2, \ldots, N$) outside of $G_N$ to $G_N$, train an SVM on feature set $G_N + \{g_i\}$, update $c$, and calculate the measures after introducing $g_i$ as follows:

$$SW(g_i) = \sum_{k=1}^{N+1} c_k = \sum_{k=1}^{N+1} w_k^2 \tag{7-4}$$

$$MW(g_i) = \max(c_k) = \max(w_k^2), k = 1, 2 \ldots N+1 \tag{7-5}$$

Here we have two strategies to choose candidates as $g_{N+1}$, corresponding to measures $SW$ and $MW$, respectively. The candidate set is denoted as $C$. The first strategy is to pick up the feature with the minimum $SW$ into $C$; and the second one is according to the minimum $MW$.

$$g_j \in C \mid SW(g_j) = \min(SW) \tag{7-6}$$

$$g_j \in C \mid MW(g_j) = \min(MW) \tag{7-7}$$

Whether set $C$ consists of multiple candidates or a single candidate, only one feature will be chosen as $g_{N+1}$. We call the support vector based recursive feature addition according to Minimum $SW$ in (7-6) and with Minimum SC in (7-3) MSW-MSC. Similarly we call the support vector based feature addition according to Minimum MW in (7-7) and with Minimum SC in (7-3) MMW-MSC.

## 7.4 Experiments and Results

### 7.4.1 Materials

*Application 1*: The role of genes and environments in the link between important health conditions: Periodontal Disease (PD) and Cardiovascular Disease (CVD). Cardiovascular disease is the number one cause of death and disability in the western world. Almost 1 million Americans die of CVD each year, which adds up to 42% of all deaths. Numerous clinical and epidemiological studies have shown a consistent association between PD and CVD and the link between these two diseases may be the result of common environmental exposures and potential genes that may regulate the individual response to these exposures. The identification of SNPs that influence the risk of diseases through interactions with other SNPs and environmental factors remains a statistical and computational challenge.

Our Myocardial Infarction Case & Control (MICC) data set is a result of a population based study. The sample included residents of Erie and Niagara counties in New York State and all were in age group 35 to 69 years. There were 614 white male patients with Myocardial Infarction matched with 614 control males (without CVD) by age (+/- 5 year) and smoking habits; 206 white pre and postmenopausal females with MI matched with 412 control females (without CVD) by age (+/- 5 year), menopausal status, years since menopause (+/- 2 year), and smoking habits. Diabetics were excluded. The features in the data set included 29 environmental variables, such as smoking status, menopausal status, blood pressure, blood cholesterol, body mass index, drinking status, etc. and 2 protein variables (ACHMN and CALMEA) that were known to be related to periodontal disease. Selection of genetic variables was based on the well known Seattle web site (http://pga.mbt.washington.edu/) using candidate approach, which included 31 SNPs in 9 genes as follows: IL 1 beta gene: rs1143634, rs16944, rs3917354, rs3917356; IL 6 gene: rs2069825, rs1818879, rs1548216, rs1800795; MMP3: rs522616, rs595840, rs602128, rs680753; TF: rs1324214, rs1361600, rs3354, rs391763. The original MICC data set contained some missing data. In our experiments, we filtered out the missing data and the associated observations. This data set was mainly used to evaluate the SNP-environment and variable-disease associations, especially the effects of SNPs and environmental variables to the disease.

*Application 2:* Rheumatoid arthritis (RA) is an autoimmune disease that causes chronic inflammation of the joints, the tissue around the joints, or other organs in the body. RA affects more than two million people in the United States. 70 percent of people with RA are women. While women are two to three times more likely to get RA, men tend to have more severe symptoms. It afflicts people of all races equally. Onset usually occurs between 30 and 50 years of age. Data for this analysis was provided as part of Genetics Analysis Workshop 15. GAW15 focused on genetic factors that predispose for rheumatiod arthritis. The North American Rheumatoid Arthritis Consortium (NARAC), lead by Peter Gregersen, has provided microsatellite and SNP scans, quantitative phenotypes, and clinical measures, with additional genotype data provided by Robert Plenge and Ann Begovich. We studied the SNP case-control data named "CHR18SNP.dat" offered by NARAC. In the data file, a dense panel of 2300 SNPs was genotyped by Illumina for an approximately 10 kb region of chromosome 18q. These markers were individually genotyped on 460 cases and 460 controls. Controls were recruited from a New York City population. The objective of this study is to identify SNPs of chromosome 18 that are significantly associated with rheumatoid arthritis. The significant SNPs identified here could be used as a starting point for biologists developing genetic tests that indicate increased risk of developing rheumatoid arthritis.

### 7.4.2 Implementations and Comparison Studies

We implemented SRFA with various statistical learning classifiers (with different complexity) proposed in section 2.1. The learning classifiers for feature selections were Naive Bayes Classifier (NBC) [Pedro and Pazzani, 1997], Nearest Mean Scaled Classifier (NMSC) [Heijden *et al*., 2004] and Dynamic Evolving Neuro-Fuzzy Inference System (DENFIS) [Kasabov, 2002; Kasabov and Song, 2002]. We recorded them as NBC-MSC, NMSC-MSC and DENFIS-MSC. Several classifiers including NBC, NMSC, SVM, uncorrelated normal based quadratic Bayes classifier that was recorded as UDC [33] were applied to the feature sets selected by the above SRFA in order to compare the performance. Our goals are (i) to evaluate feature selection procedures and find the number of features required for the best classification accuracy; (ii) to evaluate various learning approaches; (iii) to investigate the redundancy issues in SNP data for improving the classification performance.

We also implemented and tested our SVRFA: MSW-MSC and MMW-MSC methods proposed in section 2.2. For comparison purposes, other popular methods, such as Support Vector Machine Recursive Feature Elimination (SVMRFE), logistic regression based Wald t-test and Logic regression (LOGICFS) for SNP selections and disease classifications were also implemented using the R programming language. We also applied SVM and other traditional neural network classifiers such as Levenberg-Marquardt trained feed-forward neural network classifier, back-propagation trained feed-forward neural network classifier [33] for different feature selections on two real data sets. Unfortunately, these learning classifiers didn't work well. Therefore, here we did not list their experimental results.

Cross-Validation (CV) is widely used for selecting tuning parameters and optimizing the number of selected genes in the context of building classifiers to avoid over-fitting. We split the data into training and testing samples, build the model based on training samples only and evaluate the performance on the testing samples only based on cross-validation (CV). We performed 20 runs and used 50% for training and 50% for testing for each run and compared the average testing accuracy.

### 7.4.3 Results

Fig. 7-1 displays the testing accuracies of NBC, NMSC, SVM, and UDC in the analysis of the MICC data set: 4 typical runs out of 20 experiments are shown. The legend marks the different feature selections. Fig. 7-1 indicates that, NBC-MSC and NMSC-MSC feature selections are better than MSW-MSC, MMW-MSC, and SVMRFE; T-test is the worst. The comparison shows that both support vector based feature addition and SRFA with the use of different learning classifiers, the five feature selections, (MSW-MSC, MMW-MSC, NBC-MSC, NMSC-MSC, and DENFIS-MSC) on the average, outperform the popular method SVMRFE based SNP selections for disease classification in genetic association study. This demonstrates that feature addition in general is superior to feature elimination for this particular data set. Also, on the average, especially under low feature dimension, supervised recursive feature additions (SRFA) are superior to support vector based feature selections. Regarding the classification performances of different learning classifiers, on the average, NBC, NMSC, and SVM were better than UDC.

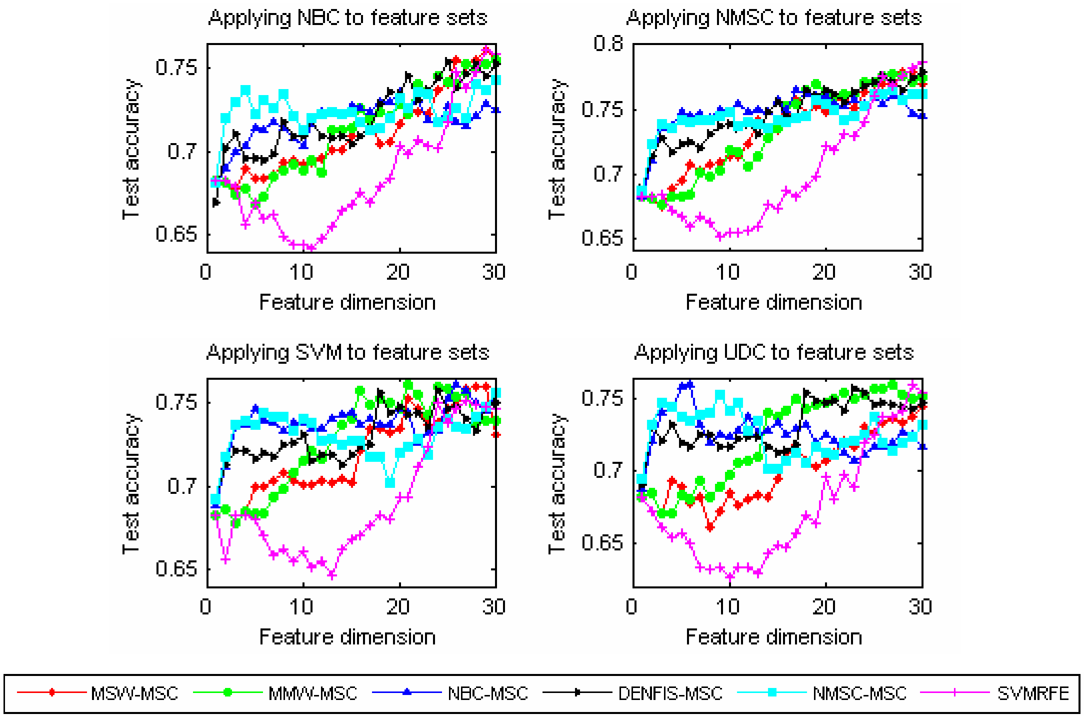


Fig. 7-1 The testing accuracies in applying NBC, NMSC, SVM, and UDC to MICC data set. The legend marks the different feature selections.

Fig. 7-2 presents the average testing accuracies on the NARAC CHR18SNP case/control data for feature dimensions 1 to 200 with the use of NBC and NMSC on the following feature selections (the legend marks the different feature selections): MSW-MSC, MMW-MSC, NBC-MSC, NMSC-MSC, SVMRFE, TTEST, and nonparametric RANKSUM. Fig. 7-2 indicates that the testing accuracies of TTEST and RANKSUM are the worst. This may be due to their selections ignoring the redundancy among SNPs, while the other five approaches (two SVRFA and three SRFA) using MSC with Spearmen Correlation Coefficients don't. MSC combined with RFA helps to improve the classification accuracy.

We noticed that as the number of features increases, the performance of the complex model, such as SVMRFE increases while simpler models stay at the same level. The reason behind this may be due to the fact that these models may detect the epistatic effects (gene-gene interactions), those that do not exhibit statistically significant marginal effects. The detection of higher dimensions of many epistatic effects requires even more complex models. In contrast, when lower levels of LD are observed at given loci, a larger number of SNPs are required to predict disease status, such as in the NARAC CHR18SNP data set. Overall, the testing accuracies of NMSC-MSC are the best, followed by NBC-MSC, MMW-MSC, MSW-MSC and SVMRFE; TTEST and RANKSUM are the worst. Comparing NBC to NMSC, on the average, the performance of NMSC is superior to NBC. Figures 7-1 and 7-2 also manifest that the classification techniques are strictly paired up with feature selections. With the use of NBC, the performance of NMSC-MSC is not so good, but with the use of NMSC, the feature selection NMSC-MSC performed the best.

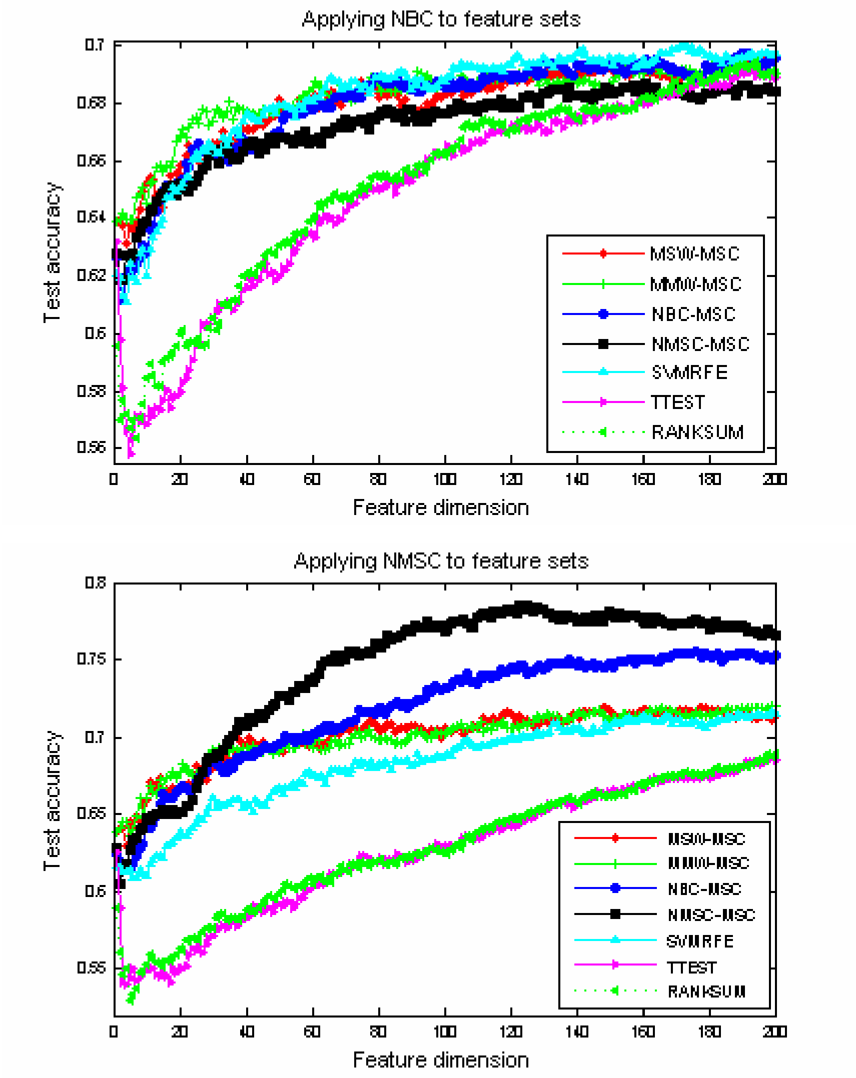


Fig. 7-2 The testing accuracies in applying NBC and NMSC to NARAC CHR18SNP data set. The legend marks the different feature selections.

Tables 7-1 and 7-2 list the testing accuracies and the standard errors associated with the highest training accuracies for given classifiers (NMSC, NBC, SVM, UDC) under different feature selections (two SVRFA: MSW-MSC, MMW-MSC; three SRFA: NBC-MSC, NMSC-MSC, DENFIS-MSC; three popular approaches: SVMRFE, Logistic-Wald-t, LOGICFS) for the MICC data set and NARAC CHR18SNP, respectively. In

Table 7-1, the testing accuracies of LOGICFS were obtained from the 31 SNPs only in the MICC data set. Table 2 indicates that, supervised learning based feature selection NMSC-MSC with the use of NMSC outperforms others, followed by NBC-MSC with the use of NMSC. Generally, support vector based feature selections are superior to LOGICFS, and LOGICFS is better than the feature selections based on parametric and non-parametric tests. Regarding support vector based feature selection, on the average, MMW-MSC outperformed MSW-MSC and SVMRFE.

Table 7-1 Testing accuracies associated with the highest training accuracies under different feature selections for the MICC data set

| Feature Selection | Testing accuracy (mean value ± standard deviation, %) | | | |
|---|---|---|---|---|
| | NMSC | NBC | SVM | UDC |
| MSW-MSC | 76.0 ± 3.4 | 75.1 ± 3.0 | 73.1 ± 4.5 | 73.6 ± 2.9 |
| **MMW-MSC** | **77.4 ± 2.9** | 75.9 ± 3.0 | 74.4 ± 2.3 | 74.8 ± 4.6 |
| NBC-MSC | 75.1 ± 3.1 | 73.2 ± 2.4 | 74.2 ± 4.1 | 75.2 ± 2.6 |
| NMSC-MSC | 75.0 ± 4.5 | 75.0 ± 2.9 | 74.0 ± 3.7 | 72.7 ± 3.9 |
| **DENFIS-MSC** | 76.9± 3.2 | 74.2 ± 3.4 | **74.9 ± 4.4** | 75.6 ± 2.8 |
| SVMRFE | 77.0 ± 4.2 | 73.9 ± 2.7 | 73.1 ± 4.0 | 74.4 ± 3.2 |
| T-TEST | 75.6 ± 2.6 | **76.4 ± 3.0** | 74.5 ± 3.1 | 75.9 ± 3.6 |
| LOGICFS | 54.4 ± 1.5 | | | |

Table 7-2 Testing accuracies associated with the highest training accuracies under different feature selections for the NARAC CHR18SNP data set

| Feature Selection | Testing accuracy (mean value ± standard deviation, %) | |
|---|---|---|
| | **NMSC** | NBC |
| MSW-MSC | 71.3 ± 0.7 | 68.5 ± 0.7 |
| **MMW-MSC** | 71.4 ± 0.4 | **69.3 ± 0.3** |
| NBC-MSC | 74.3 ± 0.6 | 68.3 ± 0.7 |
| **NMSC-MSC** | **77.7 ± 0.7** | 67.7 ± 0.3 |
| SVMRFE | 67.8 ± 0.8 | 68.3 ± 0.8 |
| T-TEST | 65.4 ± 0.5 | 66.1 ± 0.8 |
| LOGICFS | 67.1 ± 2.1 | |

## 7.5 Discussion

Exploiting information redundancy due to associations between single nucleotide polymorphism markers potentially reduces the efforts in terms of time and cost for studies since currently it is still too expensive to genotype all available SNPs across the human genome. For economic and quick diagnostic, we need advanced approaches to mine the minimum SNPs with the highest prediction accuracy for complex diseases. In this chapter we propose several new statistical learning algorithms, including SRFA and SVRFA to deal with the redundancy in the highly correlated SNP data for finding the set of SNPs enabling the most efficient classification of individuals in disease risk, which is one of the ultimate goals of human genomic research. We compared our proposed approaches with various settings (learning classifier with different complexity) to some popular methods for SNP-disease association study to see the improvement made by the proposed methods.

Compared to the well known feature selection methods SVMRFE and LOGICFS, our methods gained higher testing accuracy on the average. When SRFA is compared to two learning classifiers (NMSC-MSC and NBC-MSC), on the average, NMSC-MSC is better. Regarding SVRFA of MSW-MSC, MMW-MSC, and SVMRFE, our proposed MMW-MSC is the best. Also, on the average, SRFA performed better than SVRFA. Our study showed that using MSC for reducing the redundancy does not decrease the classification accuracy; but instead MSC combined with SRFA helps to improve the classification accuracy.

The training model is an import factor in the evaluation of the testing accuracy. In our experiments, the training with the use of DENFIS and other neural network classifiers always achieve very high training accuracy, but the testing accuracy is not so good. The occurrence of the over-fitting problem is probably related to the relatively small sample size, since complex models, such as DENFIS, almost always require large sample size to elicit their effects. While the complexity of the model increases in order to achieve higher training accuracy, the requirement for more training sample also increases. If the sample is not large enough, the relation and model mined from the training samples are not suitable for testing, and as a result over-fitting happens. This is the reason that complex models fit training samples very well, but not necessarily fit the testing samples.

Another point worthy of mentioning is that the learning classifier and feature selection are strictly paired in our models. For instance, NMSC-MSC with the use of NMSC was the best in the experiments on NARAC CHR18SNP, but NMSC-MSC with the use of NBC was not so good. The issue of environmental variables also requires discussion. With the inclusion of environmental variables in the MICC data we greatly improved the prediction and classification performances. For instance, LOGICFS only achieved 54.4%+/-1.5% correct classification rate on the testing data without the environmental variables. Also, SRFA provided a low (<60%) correct classification rate on the testing data when only using the SNPs, but a higher (>73%) correct classification rate after including the environmental variables as well. This confirms that in today's common, complex diseases, genetic and environmental variables together cause the disease and that information in necessary on both for high quality predictions and classifications.

Additionally, when SVM was applied to the feature sets extracted from the NARAC CHR18SNP genotype data, the classification performance was pretty poor. However, SVM worked well on the feature sets extracted from the MICC data. NARAC CHR18SNP consists of categorical SNP data only, while the MICC data set consists of many environmental variables of which most follow continuous distributions and have important impact on the classification. As a result, the classification with the use of support vector machines on NARAC CHR18SNP is not so good.

Our study shows that, if high level LD occurred in the population that can be captured by the classification models, only one, two or at most five SNPs would be enough to obtain a good predictive capacity. In the MICC data regions were pre-selected with high level LD using candidate gene approaches. After applying our methods, it was evident that with 3-5 variables we can achieve at least 79% classification accuracy (Fig. 7-1). On the other hand SVMRFE may capture some lower level LD and hence when the number of SNPs increases to 11-15, it achieved similar accuracies to our SRFA. In this case the simple classifier combined with our SRFA, such as NMSC-MSC or NBC-MSC is sufficient and performs better than complex models, such as SVMRFE, DENFIS, or LOGICFS. In contrast, when lower levels of LD are observed at given loci, a larger number of SNPs are required to predict disease status, such as in the case of the NARAC CHR18SNP data set. This demonstrates that the classification accuracy can be improved if prior knowledge, such high LD regions are utilized in the selections. Therefore, finding high level LD with our SRFA may directly reduce the cost of genotyping. Further investigation of whether there is power reduction compared to the selected SNPs with direct assays of all common SNPs will be conducted.

# BIBLIOGRAPHY

1. Allison, D B (2005). *DNA Microarrays and Related Genomic Techniques: Design, Analysis, and Interpretation of Experiments.* Chapman & Hall/CRC.

2. Alon,U *et al*. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proc. Natl. Acad.. Sci. USA, Cell Biology*, **96**: 6745-6750.

3. Anderson, E C and Novembre, J (2003). Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics*, **73**:336-354.

4. Avcibas, I; Memon, N and Sankur, B (2003). Steganalysis using image quality metrics. *IEEE trans. Image Processing*, **12**(2):221- 229.

5. Beckmann, L; Thomas, D C; Fischer, C; Chang-Claude, J (2005). Haplotype sharing analysis using Mantel statistics. *Human Heredity*, **59**:67-78.

6. Bo, T and Jonassen, I (2002). New feature subset selection procedures for classification of expression profiles, *Genome Biology*, **3**(4): research0017.

7. Breiman, L (2001). Random Forests. *Machine Learning* 2001, **45**:5-32.

8. Breiman, L; Friedman, J H; Olshen R A; Stone CJ (1984). *Classification and Regression Tress*, Wadsworth, Belmont.

9. Brookes, A J (1999). Review: The essence of SNPs, *Gene*, **234**(2):177-186.

10. Cardon, LR and Bell, J I (2001). Association study designs for complex diseases. *Nat. Rev. Genet.*, **2**:91–99.

11. Carlson, CS, Eberle, MA, Rieder, MJ, Yi, Q, Kruglyak, L, Nickerson, DA (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *American Journal of Human Genetics*, **74**:106-120.

12. Cores, C and Vapnik, V N (1995) Support Vector Networks. *Machine Learning,* **20**:273-297.

13. Derek Upham, http://www.nic.funet.fi/pub/crypt/steganography/.

14. Diaz-Uriarte, R and de Andres, S A (2006). Gene Selection and Classification of Microarray Data Using Random Forest*, BMC Bioinformatics***,** **7**:3.

15. Duda, R; Hart, P and Stork, D (2001). *Pattern Classification*, 2nd edition, John Wiley and Sons, New York.

16. Ekins, R and Chu, F W (1999). Microarrays: their origins and applications. *Trends Biotechnol*., **17**(6):217-218.

17. Fridrich, J (2004). Feature-Based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes, J. Fridrich (ed), 6th Information Hiding Workshop, *Lecture Notes in Computer Science,* vol. 3200, pp. 67-81.

18. Fridrich, J; Goljan, M and Du, R (2001). Detecting LSB steganography in color and gray-scale image, *IEEE Multimedia* **8**(4): 22-28.

19. Fridrich, J; Goljan, M and Hogeam, D (2003). Steganalysis of JPEG Images: Breaking the F5 Algorithm, Information Hiding: 5th Information Hiding Workshop, *Lecture Notes in Computer Science*, vol. 2578, pp. 310-323.

20. Fridrich, J; Soukal, D and Goljan, M (2005). Maximum Likelihood Estimation of Length of Secret Message Embedding using ±K Steganography in Spatial Domain, *Proc. Of SPIE - Secruity, Steganography, and Watermarking of Multimedia Contents, VII*, E. Delp III, P. Wong (eds.), vol. 5681, pp. 595-606.

21. Friedman, J; Hastie, T and Tibshirani, R (2000). Additive Logistic Regression: A Statistical View of Boosting**.** *The Annals of Statistics*, **38**(2): 337–374.

22. Golub, T *et al*. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression, *Science*, **286:** 531-537.

23. Gopalakrishnan, S and Qin, Z S (2006). TagSNP Selection Based on Pairwise LD Criterion and Power Analysis in Association Studies. *Pacific Sym. Biocomputing* 11:511-522.

24. Guyon, I; Weston, J; Barnhill, S and Vapnik, V (2002). Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning,* **46**(1-3):389–422.

25. Halldorsson, B V; Bafna, V; Lippert, R; Schwartz, R; De La Vega, F M; Clark, A G; Istrail, S (2004). Optimal haplotype block-free selection of tagging SNPs for genomewide association studies. *Genome Research,* **14**:1633-1640.

26. Hampe J, Schreiber S, Krawczak M (2003). Entropy-based SNP selection for genetic association studies. *Hum. Genet*., **114**:36-43.

27. Hand, D J and Heard, N A (2005). Finding Groups in Gene Expression Data. *J. Biomed Biotechnol.,* **2**: 215-225.

28. Harmsen, J J and Pearlman, W A (2003). Steganalysis of additive-noise modelable information hiding, *Proc. of SPIE: Security and Watermarking of Multimedia Contents V*, Edward J. Delp III, Ping W. Wong, Editors, vol. 5020, pp. 131-142.

29. Harmsen, J J and Pearlman, W A (2004). Kernel Fisher Discriminant for Steganalysis of JPEG Hiding Methods, *Proc. of SPIE, Security, Steganography, and Watermarking of Multimedia Contents* VI, vol. 5306, pp.13-22.

30. He, J and Zelikovsky, A (2006). MLR-tagging informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics* **22**(20): 2558-2561.

31. Heijden, F; Duin, R; Ridder, D; Tax, D (2004). *Classification, Parameter Estimation and State Estimation*. John Wiley.

32. Huang, J and Mumford, D (1999). Statistics of Natural Images and Models. *Proc. of Computer Vision and Pattern Recognition* (CVPR'99), vol.1, DOI: 10.1109/CVPR.1999.786990.

33. Inza, I; Sierra, B; Blanco, R and Larranaga, P (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent and Fuzzy Systems,* **12**(1): 25-33.

34. Kasabov, N (2002). *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines*. London-New York, Springer-Verlag.

35. Kasabov, N and Song, Q (2002). DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction, *IEEE Trans. Fuzzy Systems,* **10**(2):144-154.

36. Katzenbeisser, S and Petitcolas, F (2000). *Information Hiding Techniques for steganography and Digital Watermarking*. Artech House Books.

37. Kelley J (2001). Terror groups hide behind Web encryption, *USA TODAY*, http://www.usatoday.com/tech/news/2001-02-05-binladen.htm.

38. Ker, A (2005). Steganalysis of LSB Matching in Grayscale Images, *IEEE Signal Processing Letters* **12** (6): 441-444.

39. Kooperberg, C; Ruczinski, I; LeBlanc, M and Hsu, L (2001). Sequence Analysis Using Logic Regression. *Genetic Epidemiology* **21**:626-631.

40. Krishnapuram, B; Carin, L; Figueiredo, M and Hartemink, A. (2005). Sparse multinomial logistic regression: fast algorithms, and generalization bounds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **27**(6): 957-968.

41. Kurak, C and McHugh, J (1992). A Cautionary Note on Image Downgrading, *Proc. of 8th Annual Computer Security Applications Conference*, pp. 153-159.

42. Lander *et al*. (1999). The Chipping Forecast, *Nature Genetics* (*Supplement*), **21**(1):1-60.

43. Levin, AM *et al*. (2005). A model-based scan statistics for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics,* **21**:2867–2874

44. Liang, Y and Kelemen, A (2005). Temporal Gene Expression Classification with Regularised Neural Network. *International Journal of Bioinformatics Research and Applications*, **1**(4): 399-413.

45. Lin, Z and Altman, R B (2004). Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics,* **75**:850-861.

46. Liu, Q and Sung, A H (2006). Recursive Feature Addition for Gene Selection. *Proc. of 19th International Joint Conference on Neural Network*. pp. 2339-2346.

47. Liu, Q and Sung, A H (2007). Feature Mining and Nuero-Fuzzy Inference System for Steganalysis of LSB Matching Steganography in Grayscale Images. *Proc. of 20th International Joint Conference on Artificial Intelligence*, pp. 2808-2813.

48. Liu, Q; Sung, A H; Chen, Z and Xu, J (2007). Feature Mining and Pattern Classification for LSB Matching Steganography in Grayscale Images. *Pattern Recognition*, doi: 10.1016/j.patcog.2007.06.005.

49. Liu, Q; Sung, A H and Ribeiro, B M (2005). Statistical Correlations and Machine Learning for Steganalysis, *Adaptive and Natural Computing Algorithms*, Ribeiro *et al.* (eds.), Springer-Wien NewYork, pp. 437-440.

50. Liu, Q; Sung, A H and Xu, J (2005). Detection of Hidden Data in Digital Media, *Proc. of 2nd International Conference on Intelligent Knowledge Systems*, pp.107-111.

51. Liu, Q; Sung, A H; Xu, J and Ribeiro, B M (2006). Image Complexity and Feature Extraction for Steganalysis of LSB Matching Steganography. *Proc. of 18th International Conference on Pattern Recognition*, vol. 2, pp. 267-270.

52. Liu, Q; Sung, A H; Xu, J; Venkataramana, V (2006). Detect JPEG Steganography Using Polynomial Fitting. *Proc. of 16th Artificial Neural Networks in Engineering*, ASME Press, 2006**,** pp. 547-556.

53. Long, A; Mangalam, H; Chan, B; Tolleri, L; Hatfield, G and Baldi, P (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework, *J. Biol. Chem.*, **276**:19937-19944.

54. Lyu, S and Farid, H (2004). Steganalysis using Color Wavelet Statistics and One-class Support Vector Machines, in *SPIE Symposium on Electronic Imaging*, San Jose, CA, 2004.

55. Lyu, S and Farid, H (2005). How Realistic is Photorealistic, *IEEE Trans. on Signal Processing*, **53**(2): 845-850.

56. Mallat, S (1999). *A Wavelet Tour of Signal Processing*. Academic Press.

57. Mannila, H; Koivisto, M; Perola, M; Varilo, T; Hennah, W; Ekelund, J; Lukk, M; Peltonen, L; Ukkonen, E (2003). Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *American Journal of Human Genetics,* **73**:86–94.

58. Marvel, L M; Boncelet, C G; Jr. Retter, C T (1999). Spread Spectrum Image Steganography, *IEEE Trans. Image Processing*, **8**(8):1075-1083.

59. Monari, G and Dreyfus, G (2000). Withdrawing an example from the training set: an analytic estimation of its effect on a nonlinear parameterized model, *Neurocomputing Letters*, **35**:195-201.

60. Motulsky, H (1995). *Intuitive Biostatistics*, Oxford University Press.

61. Moulin, P and Liu, J (1999). Analysis of Multiresolution Image Denoising Schemes using Generalized Gaussian and Complexity priors, *IEEE Trans. Inform. Theory*, **45**(3): 909–919.

62. Muller, U R and Nicolau, D V (Eds.) (2005). *Microarray Technology and Its Applications*. Springer.

63. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001), On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology*, **8**(1):37-52.

64. Pedro D, Pazzani M (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, **29**:103–137.

65. Petitcolas, F A P; Anderson, R J and Kuhn, M G (1999). Information Hiding – A Survey, *Proc. of the IEEE, special issue on protection of multimedia contents*, **87**(7): 1062-1078.

66. Pomeroy, S L *et al*. (2002). Prediction of Central Nervous System Embryonal Tumor Outcome based on Gene Expression, *Letters to Nature, Nature,* **415**: 436-442.

67. Provos, N (2001). Defending against Statistical Steganalysis. *Proc. of the 10th USENIX Security Symposium*, pp.323-335.

68. Provos, N and Honeyman, P (2003). Hide and Seek: An Introduction to Steganography, *IEEE Security & Privacy*, **1**(3): 32- 44.

69. Pvlidis, P and Noble, W S (2001). Analysis of strain and regional variation in gene expression in mouse brain, *Genome Biology*, **2**(10): research0042.1-0042.15.

70. Qin, Z S (2006). Clustering Microarray Gene Expression Data using Weighted Chinese Restaurant Process, *Bioinformatics*, **22**(16): 1988-1997.

71. Quackenbush, J (2001) Computational Analysis of Microarray Data, *Nature Rev.Genteic*, **2**: 418-427.

72. Ramkumar, M; Akansu, A and Alatan, A (1999). A Robust Data Hiding Scheme for Digital Images Using DFT, *Proc. of IEEE International Conference on Image Processing* 1999, vol 2, pp 211-215.

73. Rencher, A C (2002). *Methods of Multivariate Analysis,* Wiley-Interscience; 2nd edition.

74. Risch, N J (2000). Searching for genetic determinants in the new millennium. *Nature,* **405**:847–856.

75. Rivals, I; Personnaz, L (2003). MLPs (Mono-Layer Polynomials and Multi-Layer Perceptrons) for Nonlinear Modeling, *Journal of Machine Learning Research*, **3**:1383-1398.

76. Schaid, D J; Rowland, C M; Tines, D E; Jacobson, R M; Poland, G A (2002). Score test for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics,* **70**:425—443.

77. Schlesinger, M and Hlavac, V (2002). *Ten Lectures on Statistical and Structural Pattern Recognition*, Kluwer Academic Publishers.

78. Schwender, H and Ickstadt, K (2006). Identification of SNP Interactions Using Logic Regression, http://www.sfb475.uni-dortmund.de/berichte/tr31-06.pdf, accessed on Oct.-31-2006.

79. Sebastiani, P; Lazarus, R; Weiss, S T; Lunkel, L M; Kohane, I S; Romani, M F (2003). Minimal haplotype tagging, *Proc. Natl. Acad. Sci.,* **100**:9900-9905.

80. Segal, E; Friedman, N; Kaminski, N; Regev, A and Koller, D (2005). From Signatures to Models: Understanding Cancer Using Microarrays, *Nature Genetics*, **37**:S38-45.

81. Sha, N; Tadesse, M G; Vannucci, M (2006). Bayesian Variable Selection for the Analysis of Microarray Data with Censored Outcomes, *Bioinformatics,* **22**(18):2262-2268.

82. Sharifi, K and Leon-Garcia, A (1995). Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video, *IEEE Trans. Circuits Syst. Video Technol*., **5**(1): 52–56.

83. Sharp, T (2001). An Implementation of Key-Based Digital Signal Steganography, in I. Moskowitz (ed.): Information Hiding. 4th International Workshop, *Lecture Notes in Computer Science*, vol. 2137, pp. 13–26. Springer-Verlag.

84. Shipp, M. *et al*. (2002) Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning, *Nature Medicine,* **8**(1):68-74.

85. Simon, R M; Korn, E L; McShane, L M; Radmacher, M D; Wright, G W and Zhao, Y (2003). *Design and Analysis of DNA Microarray Investigations*, Springer.

86. Singh, D *et al*. (2002) Gene Expression Correlates of Clinical Prostate Cancer Behavior, *Cancer Cell,* **1**(2): 227-235.

87. Song, K; Elston, R C (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat. Med.,* **25**:105-126.

88. Stram, D O; Haiman, C A; Hirschhorn, J N; Altshuler, D; Kolonel, L N; Henderson, B E; Pike, M C (2003). Choosing haplotype-tagging SNPs based on unphased genotype data using preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum. Hered.,* **55**:27-36.

89. Takagi, T and Sugeno, M (1985). Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-15, pp. 116-132.

90. Tan, P; Steinbach, M and Kumar, V (2005). *Introduction to Data Mining*, Addison-Wesley, pp. 76-79.

91. Tang, E K; Suganthan, P N and Yao, X (2006). Gene Selection algorithms for Microarray Data Based on Least Square Support Vector Machine, *BMC Bioinformatics*, **7**:95. doi: 10.1186/1471-2105-7-95.

92. Taylor, J and Cristianini, N (2004). *Kernel Methods for Pattern Analysis,* Cambridge University Press.

93. The International HapMap Consortium (2003). The International HapMap Project. *Nature*, **426**:789–796.

94. The International HapMap Consortium (2004). Integrating ethics and science in the International HapMap Project. *Nat Rev Genet*, **5**:467–475.

95. The International HapMap Consortium (2005). Haplotype map of the human genome. *Nature*, **437**:1299-1320.

96. Tibshirani, R (1996). Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc B.,* **58**(1): 267-288.

97. Tibshirani, R. (1997). The lasso method for variable selection in the Cox model, *Statistics in Medicine,* **16**:385-395.

98. Tjaden, B (2006). An Approach for Clustering Gene Expression Data with Error Information, *BMC Bionformatics*, **7**:17, doi: 10.1186/1471-2105-7-17.

99. Van, L J *et al*. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature, Nature,* **415**: 530-536.

100. Vapnik, V N (1998). *Statistical Learning Theory,* John Wiley.

101. Vapnik V N (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

102. Wainwright, M and Simoncelli, E (2000). Scale Mixtures of Gaussians and the Statistics of Natural Images, in: *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller(Eds.), Cambridge, MA: MIT Press, vol. 12, pp. 855–861.

103. Webb, A (2002). *Statistical Pattern Recognition*, John Wiley & Sons, New York.

104. Westfeld, A (2001). High Capacity Despite Better Steganalysis (F5–A Steganographic Algorithm), Proc. of 4th Information Hiding Workshop, *Lecture Notes in Computer Science*, vol. 2137, pp. 289–302.

105. Westfeld, A and Pfitzmann, A (2000). Attacks on Steganographic Systems. *LNCS*, vol.1768, pp. 61-75.

106. Winkler, G (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Springer.

107. Witte, J S and Fijal, B A (2001). Introduction: Analysis of Sequence Data and Population Structure. *Genetic Epidemiology*, **21**:600-601.

108. Wouwer, G; Scheunders, P and Dyck, D (1999). Statistical Texture Characterization from Discrete Wavelet Representations, *IEEE Trans. Image Processing*, **8**(4): 592–598.

109. Xu, J; Sung, A H; Shi, P and Liu, Q (2003). Text Steganography using Wavelet Transform, *Proc. of 7th IASTED International Conference on Internet and Multimedia Systems and Applications*, pp. 473-478.

110. Xu, J; Sung, A H; Shi, P and Liu, Q (2004). JPEG Compression Immune Steganography Using Wavelet Transform, *Proc. of International Conference on Information Technology: Coding and Computing 2004 (ITCC 2004),* vol. 2, pp. 704 - 708.

111. Yu *et al.* (2004). A Secure Steganographic Scheme against Statistical Analyses. *Lecture Notes in Computer Science*, pp. 497-507.

112. Yu, J and Chen, X (2005). Bayesian Neural Network Approaches to Ovarian Cancer Identification from High-resolution Mass Spectrometry Data, *Bioinformatics,* vol. 21 (suppl-1), pp. i487-i494.

113. Zhang, K and Jin, L (2003). HaploBlockFinder: Haplotype block analysis. *Bioinformatics*, **19**:1300-1301.

114. Zhang, T and Ping, X (2003). A Fast and Effective Steganalytic Technique against JSteg-like Algorithms, *Proc. 8th ACM Symp. Applied Computing*, ACM Press, pp. 307 - 311.

115. Zhou, X and Mao, K Z (2005). LS Bound Based Gene Selection for DNA Microarray Data, *Bioinformatics* **21**(8):1559-1564.